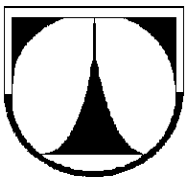


TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

Katedra elektroniky a zpracování signálů

DIPLOMOVÁ PRÁCE

Simulace rušivých vlivů přenosové cesty akustického signálu při rozpoznávání řeči po telefonu

The simulation of disturbing effects of the telephone line at speech recognition over the telephone

Liberec 2003 **Pavel Semenec**

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

Katedra elektroniky a zpracování signálů Akademický rok: 2002/2003

ZADÁNÍ DIPLOMOVÉ PRÁCE

pro: **Pavla SEMENCE**

studijní program: 2612 M – Elektrotechnika a informatika

obor: 2612 T – Automatické řízení a inženýrská informatika

Vedoucí katedry Vám ve smyslu zákona o vysokých školách č.111/1998 Sb. určuje tuto diplomovou práci:

Název tématu:

Simulace rušivých vlivů přenosové cesty akustického signálu při rozpoznávání řeči po telefonu

Zásady pro vypracování:

1. Seznamte se s problematikou rozpoznávání řeči, především s rozpoznáváním izolovaných slov a krátkých frází pomocí skrytých markovských modelů.
2. Prostudujte celý rozpoznávací řetězec a seznamte se s problematikou rozpoznávání řečového signálu získaného z telefonní linky.
3. Navrhněte metody simulující rušivé vlivy telefonní přenosové cesty.
4. Navržené metody realizujte v kompaktním programu s možností vnější konfigurace.
5. Navržený software otestujte v rozpoznávacích experimentech na rozsáhlejší databázi a získejte optimální konfiguraci pro simulaci rušivých vlivů.

Rozsah grafických prací: dle potřeby dokumentace

Rozsah průvodní zprávy: cca 40 až 50 stran

Seznam odborné literatury:

[1] Psutka J.: Komunikace s počítačem mluvenou řečí. Academia, Praha, 1995

[2] Nouza J. (editor): Počítačové zpracování řeči. Sborník článků, Liberec, 2001

[3] HOLADA M., NOUZA J.: Searching for Methods and Parameters for More Reliable Recognition of Telephone Speech. Proc. of Radioelektronika'98, Brno, April 1998, p.220-223

[4] Huang X., Acero A., Hon H.W.: Spoken Language Processing. Prentice Hall 2001.

[5] články ze sborníků konferencí Eurospeech a ICSLP

Vedoucí diplomové práce: Ing. Miroslav Holada

Konzultant:

Zadání diplomové práce: 21.10.2002

Termín odevzdání diplomové práce: 23.05.2003

.....

Vedoucí katedry

.....

Děkan

V Liberci dne 21. 10. 2002

Anotace

Cílem diplomové práce bylo navrhnout metody simulace rušivých vlivů telefonní linky pro úpravu mikrofonní trénovací množiny rozpoznávače řeči, které by vedly ke zlepšení úspěšnosti rozpoznávání řeči po telefonní lince bez nutnosti pořizování telefonní trénovací množiny. Dalším úkolem byla realizace těchto metod v kompaktním programu s možností vnější konfigurace, nalezení optimálních parametrů simulace a ověření funkčnosti a úspěšnosti metod na experimentech s rozpoznávačem izolovaných slov.

Byly navrženy dvě rozdílné metody simulace telefonní linky, první pracuje s bankou filtrů, druhá s identifikovaným modelem. Rozpoznávací experimenty potvrdily, že obě metody jsou funkční, v prvním případě došlo k poměrnému zlepšení rozpoznávacího skóre o 52,22 %, v druhém případě o 33,72 %.

Abstract

The aim of the diploma thesis was to design simulation methods of disturbing effects of the telephone line to modify a speech recognizer microphone training set which are made to higher correctness at speech recognition over the phone with no need of a telephone training set. Next job was to realize these methods as a compact program with an external configuration possibility, to search the optimal simulation parameters and to verify the methods functionality on recognition experiments with isolated words.

Two different methods of the telephone line simulation were designed, the former method uses a bank of filters and the later method uses an identified model. Experiments confirmed that both methods are functional, in the first case the relative improvement of correctness was 52,22 %, in second case it was 33,72 %.

Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 o právu autorském, zejména § 60 (školní dílo).

Beru na vědomí, že TUL má právo na uzavření licenční smlouvy o užití mé DP a prohlašuji, že **s o u h l a s í m** s případným užitím mé diplomové práce (prodej, zapůjčení apod.).

Jsem si vědom toho, že užít své diplomové práce či poskytnout licenci k jejímu využití mohu jen se souhlasem TUL, která má právo ode mne požadovat přiměřený příspěvek na úhradu nákladů, vynaložených univerzitou na vytvoření díla (až do jejich skutečné výše).

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím diplomové práce a konzultantem.

Datum

Podpis

1 Obsah

1 Obsah 5

2 Předmluva 7

3 Úvod 8

4 Teorie číslicového zpracování akustického signálu 9

4.1 Digitalizace spojitého akustického signálu 9

4.2 Číslicová filtrace akustického signálu 11

5 Teorie automatického rozpoznávání řeči 14

5.1 Parametrizace řečového signálu 14

5.2 Příznaky určené v časové oblasti 16

5.3 Kepstrální příznaky 17

5.3.1 Kepstrální příznaky LPCC (Linear Predictive Cepstral Coefficients) 18

5.3.2 Kepstrální příznaky MFCC (Mel-frequency Cepstral Coefficients) 19

5.4 Rozpoznávače řeči 21

5.4.1 Klasifikace slov pomocí referencí, algoritmus DTW 21

5.4.2 Klasifikace slov pomocí skrytých markovských modelů 23

6 Realizace simulátoru telefonní linky 28

6.1 Metoda první - banka filtrů 28

6.1.1 Pořízení zkušebních telefonních nahrávek 28

6.1.2 Analýza zkušebních telefonních nahrávek 29

6.1.3 Návrh softwaru 32

6.1.4 Popis konfigurace softwaru 35

6.2 Metoda druhá – identifikovaný model 36

6.2.1 Pořízení telefonních nahrávek pro identifikaci 36

6.2.2 Odhad parametrů modelu 37

6.2.3 Návrh softwaru 39

6.2.4 Popis konfigurace softwaru 40

7 Rozpoznávací experimenty 42

7.1 Pořízení nahrávek trénovací a testovací množiny 43

7.2 Výsledky rozpoznávacích experimentů 44
7.2.1 Experimenty s neupravenou trénovací množinou 45
7.2.2 Experimenty s trénovací množinou upravenou bankou filtrů 46
7.2.3 Experimenty s trénovací množinou upravenou identifikovaným modelem 50
7.2.4 Experimenty s trénovací množinou upravenou kombinací obou metod 52
7.3 Zhodnocení výsledků rozpoznávacích experimentů 53
8 Závěr 55
9 Literatura 57
10 Přílohy 59
10.1 Slovník slov použitých v rozpoznávacích experimentech 59
10.2 Datová příloha na disku CD-R 60

2 Předmluva

Rád bych tohoto místa využil k poděkování všem lidem, kteří mě nejrůznějšími způsoby pomáhali a podporovali při tvorbě této diplomové práce.

V první řadě děkuji Ing. Miroslavu Holadovi, vedoucímu diplomové práce, za velmi vstřícný přístup, velkou ochotu, porozumění a za mnoho užitečných konzultací a rad.

Děkuji také Prof. Ing. Janu Nouzovi, CSc. za užitečné rady a za jeho zájem o téma této diplomové práce.

Za zapůjčení přenosného počítače, bez kterého by bylo pořizování různých mikrofonních nahrávek mnohem složitější, děkuji Ing. Mirko Šormovi.

Děkuji také všem lidem, kteří mi věnovali trochu svého času pro namlouvání nahrávek do trénovací a testovací množiny.

Nakonec bych chtěl poděkovat za obrovskou podporu nejen při tvorbě této diplomové práce, ale po celou dobu mého studia, svým rodičům a vůbec celé své rodině, do které samozřejmě počítám také svou přítelkyni Marcelu.

3 Úvod

V současné době se na poli informačních systémů začínají stále více prosazovat tzv. hlasové dialogové systémy [1], [2]. Jak z názvu vyplývá, jedná se o automatické systémy, které jsou schopny s uživatelem vést řečový dialog a na jeho základě provést nějakou akci, např. poskytnout určitou informaci. Důležitou součástí takového systému je automatický rozpoznávač řeči. Jak z dalších kapitol vyplyne, pro co nejlepší funkčnost takového rozpoznávače není důležitá pouze jeho struktura a použité algoritmy. Velkou roli v ovlivnění kvality, tedy úspěšnosti, rozpoznávání hraje také tzv. fáze trénování. V ní se rozpoznávači předkládá určitá trénovací množina, ve které jsou přítomny nahrávky lidské řeči se známým obsahem, podle kterých se učí porozumět lidské řeči. Na kvalitní trénovací množinu jsou přitom kladeny určité nároky. Měla by být dostatečně obsáhlá, zároveň by měla co možná nejlépe postihovat sociologické zastoupení potenciálních uživatelů, tedy obsahovat ve správném poměru nahrávky obou pohlaví vždy v různých věkových kategoriích. Nelze však ani opomíjet prostředí, ve kterém bude aplikace nasazena, a s ním související technickou kvalitu nahrávek, neboť tyto jevy mohou negativním způsobem ovlivňovat nebo dokonce zcela degradovat některé parametry lidského hlasu používané pro jeho rozpoznávání. Snad nejmarkantněji se problém prostředí a technické kvality akustického signálu projevuje v aplikacích pracujících po telefonní lince.

Z uvedeného je zřejmé, že vytvoření kvalitní trénovací množiny je velmi pracná, časově a mnohdy i finančně náročná činnost. Přitom pro každou další aplikaci by se měla pro dosažení kvalitnějších výsledků tato množina pořizovat znovu, na konkrétním zařízení a v prostředí, ve kterém bude aplikace užívána. To je ovšem přístup velmi neefektivní. Proto v Laboratoři počítačového zpracování řeči na Technické univerzitě v Liberci vznikla myšlenka pracovat stále jen s jednou, velmi kvalitní množinou a tuto množinu pro různé další aplikace pouze softwarově upravovat. Zmíněná množina vznikala v této laboratoři v průběhu několika let nahráváním řeči do počítače pomocí různých typů mikrofonů.

Protože hlasové dialogové systémy jsou dnes často provozovány přes telefonní linku, vznikl požadavek na vytvoření softwaru, který by byl schopen rušivé vlivy této linky simulovat. Z tohoto požadavku vzešlo také téma mé diplomové práce. Mým cílem bylo nejen vytvořit zmíněný software, ale také ověřit jeho funkčnost na řadě rozpoznávacích experimentů, tedy vyzkoušet, zda se po úpravě trénovací, původně mikrofonní, množiny zlepší výsledek rozpoznávání řeči po telefonní lince.

4 Teorie číslicového zpracování akustického signálu

4.1 Digitalizace spojitého akustického signálu

Prvním krokem nutným k číslicovému zpracování akustického signálu je jeho digitalizace, neboť kmity, kterými je akustický signál reprezentován, resp. elektrický signál získaný z kmitů mikrofonem, jsou spojité. Proces digitalizace v sobě zahrnuje vzorkování, kvantování a následné kódování signálu [1], [3].

Vzorkováním se rozumí postupný převod spojitého signálu na posloupnost čísel, diskrétních vzorků. Okamžitá hodnota vzorku se odečítá vždy po uplynutí určitého časového intervalu. Ve většině aplikací je tento interval konstantní a označuje se jako tzv. perioda vzorkování t_s. Častěji se však pracuje s pojmem vzorkovací frekvence f_s, která je rovna převrácené hodnotě periody vzorkování. Mimo hardwarového je jediným, zato však podstatným, omezením volby vzorkovací frekvence podmínka vyplývající z Shannon – Kotělnikovova (Nyquistova) vzorkovacího teorému. Ten říká, že spojitý signál lze zrekonstruovat z číslicového pouze tehdy, pokud je vzorkovací frekvence vyšší než dvojnásobek nejvyšší frekvence obsažené ve spektru vzorkovaného signálu. Nedodržením této podmínky dochází k jevu zvanému aliasing – přeložení frekvencí, který vede ke zkreslení signálu. Před samotným vzorkováním se proto spojitý signál ještě frekvenčně omezuje, přesněji řečeno filtruje, vhodnou dolní propustí, tzv. antialiasingovým filtrem. Ten bývá většinou součástí zařízení, které provádí digitalizaci, u počítače jej tedy obsahuje zvuková karta.

Dalším krokem digitalizace je kvantování a s ním úzce související kódování. Proces kvantování zajišťuje převod analogových hodnot vzorků na konečný počet hodnot číslicových, kódování určuje, v jakém formátu jsou tyto číslicové vzorky vyjádřeny. Rozšířené je kvantování lineární, kde jsou všechny kvantizační intervaly stejně velké, méně často se používá kvantování nelineární, kde se velikost intervalů mění, nejčastěji logaritmicky. Protože počet kvantizačních úrovní je konečný (zpravidla se volí jako 2^n, kde n je počet bitů reprezentujících vzorek), vždy dochází k určité chybě převodu, která je označována jako kvantizační šum. Důležitým parametrem je SNR – odstup signálu od kvantizačního šumu, při lineárním kvantování pro něj platí vztah

$$\mathrm{SNR} = 6{,}02 \cdot \mathrm{n} + 1{,}76 \quad [\mathrm{dB}] \,, \tag{4.1}$$

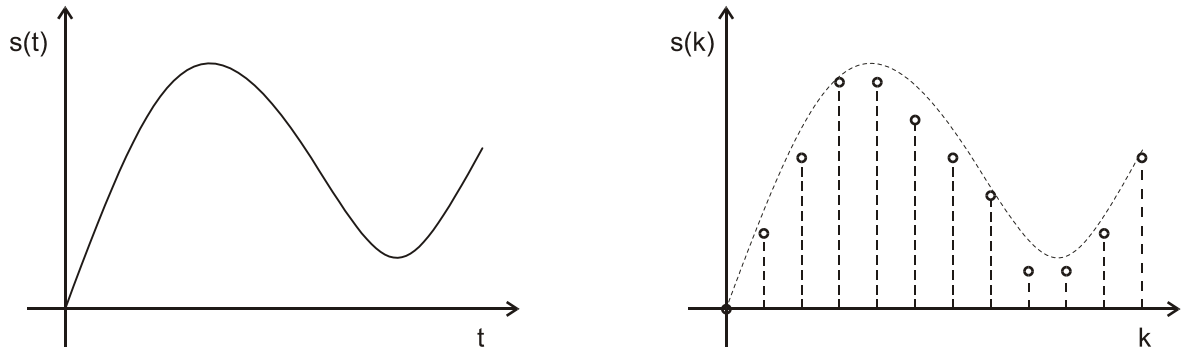
kde n je opět počet bitů reprezentujících vzorek. Z tohoto vztahu mimo jiné plyne, že SNR se s každým dalším bitem zlepšuje o 6 dB. Nejčastějším způsobem kódování je postupný

záznam číslicových hodnot jednotlivých vzorků. Pokud je toto kódování použito v kombinaci s lineárním kvantováním, hovoří se o tzv. pulzní kódové modulaci (PCM), která se používá také v nejrozšířenější verzi zvukových souborů typu WAV. Nelineární logaritmické kvantování spolu s tímto kódováním, které se nejčastěji nasazuje do telefonních aplikací, je označováno jako A-law nebo také μ-law. Existují ale také způsoby kódování, které nepracují s aktuálními hodnotami jednotlivých vzorků. Jednou z takových kódovacích technik je diferenční pulzní kódová modulace (DPCM), která zaznamenává místo aktuálních hodnot vzorků jen jejich změny oproti hodnotám předchozím. Dalšími představiteli této skupiny jsou např. adaptivní diferenční pulzní kódová modulace (ADPCM), delta modulace (DM), či adaptivní delta modulace (ADM).

Nejpoužívanější parametry digitalizace spolu s oblastí použití zobrazuje tabulka 4.1. Pro potřeby zpracování lidské řeči se jako určitý standard ustálilo kódování PCM, 16 bitů, 8 kHz, které je nejlepším kompromisem mezi potřebnou kvalitou záznamu a objemností dat. Výsledek procesu digitalizace je znázorněn na obrázku 4.1.

Tab. 4.1. Nejpoužívanější parametry digitalizace

Oblast použití	**Kódování**	**Počet bitů na vzorek**	**Vzorkovací frekvence (kHz)**
záznam hudby na CD	PCM	16	44,1
kvalitní záznam řeči	PCM	16	16
běžný záznam řeči	PCM	16	8
telefonní záznam řeči	A-law, μ-law	8	8



Obr. 4.1. Spojitý akustický signál s(t) a jeho číslicová podoba s(k) získaná digitalizací

4.2 Číslicová filtrace akustického signálu

Jednou z možností popisu číslicové soustavy je popis pomocí impulsní odezvy, která je definována jako reakce ustálené soustavy na Diracův jednotkový impuls. Číslicové filtry se podle charakteru své impulzní odezvy dělí na dva základní typy, filtry typu FIR a IIR [4], [5].

Filtr typu FIR (Finite Impulse Response) je filtrem s konečnou impulsní odezvou, jehož výstup je závislý pouze na vstupu. Diferenční rovnice takového filtru má tvar

$$y(k) = b_0 \cdot x(k) + b_1 \cdot x(k-1) + b_2 \cdot x(k-2) + ... + b_M \cdot x(k-M) \ , \tag{4.2}$$

kde y(k) je výstup filtru, x(k) je vstup filtru a M je řád filtru (délka filtru je rovna M+1). Přenosová funkce H(z) je rovna

$$H(z) = \frac{Y(z)}{X(z)} = b_0 + b_1 \cdot z^{-1} + b_2 \cdot z^{-2} + ... + b_M \cdot z^{-M} \ , \tag{4.3}$$

kde Y(z) je z-obraz výstupu, X(z) je z-obraz vstupu a z^{-1} je operátor jednotkového zpoždění.

U filtru typu IIR (Infinite Impulse Response), filtru s nekonečnou impulsní odezvou, se na výstup promítá nejen vstup, ale také minulé hodnoty výstupu. Diferenční rovnice má proto tvar

$$y(k) + a_1 \cdot y(k-1) + ... + a_N \cdot y(k-N) = b_0 \cdot x(k) + b_1 \cdot x(k-1) + ... + b_M \cdot x(k-M) \ . \tag{4.4}$$

Přenosová funkce H(z) je potom rovna

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0 + b_1 \cdot z^{-1} + b_2 \cdot z^{-2} + ... + b_M \cdot z^{-M}}{1 + a_1 \cdot z^{-1} + a_2 \cdot z^{-2} + ... + a_N \cdot z^{-N}} \ . \tag{4.5}$$

Základní vlastnosti filtrů typu FIR a IIR jsou odlišné. Jak ukazuje tabulka 4.2, každý má své výhody i nevýhody.

Tab. 4.2. Porovnání vlastností filtrů FIR a IIR

Výhody filtrů FIR oproti IIR	**Výhody filtrů IIR oproti FIR**
vždy stabilní	dostačují nižší řády filtrů
lineární fázová charakteristika	kratší skupinové zpoždění

Při znalosti impulsní odezvy h(k), resp. přenosové funkce H(z), konkrétního filtru existuje několik jednoduchých postupů k určení výstupní posloupnosti y(k) ze známé vstupní posloupnosti x(k). Výpočet v diskrétní časové oblasti lze provést např. podle vztahu

$$y(k) = \sum_{n=-\infty}^{\infty} x(n) \cdot h(k-n) = x(k) * h(k) \ , \tag{4.6}$$

kde operace označená symbolem $*$ se nazývá konvoluce. Další možností výpočtu v časové oblasti je rekurzivní výpočet s uvažováním počátečních podmínek, který se provádí postupným dosazováním do diferenční rovnice filtru pro k = 0, 1, 2, ….. . Tento postup je někdy označován jako Pierceův algoritmus.

Existuje i způsob výpočtu ve frekvenční oblasti podle předpisu

$$Y(n) = X(n) \cdot H(n) \ , \tag{4.7}$$

kde X(n) jsou frekvenční vzorky vstupního signálu určené z posloupnosti x(k) pomocí diskrétní Fourierovy transformace (DFT) a H(n) jsou frekvenční vzorky impulsní odezvy filtru. Ty se mohou určit obdobně ze vzorků h(k), nebo z frekvenční charakteristiky H(F), pro kterou platí

$$H(F) = H(z)\big|_{z=e^{j\cdot 2\cdot\pi\cdot F}} \ , \tag{4.8}$$

kde $F = \frac{f}{f_s}$ je tzv. digitální frekvence. Výstupní posloupnost y(k) se následně jednoduše určí z Y(n) aplikací zpětné diskrétní Fourierovy transformace (IDFT).

Jedním z postupů návrhu filtru typu FIR je návrh pomocí zpětné diskrétní Fourierovy transformace, často označovaný jako metoda oken [4]. Hlavní myšlenkou této metody je návrh filtrů pomocí obdélníků ve frekvenční oblasti a následný převod těchto frekvenčních obdélníků do časové oblasti pomocí IDFT. Touto operací se získá nekonečná impulsní odezva, kterou lze snadno omezit vhodným oknem na konečný počet vzorků. Velkou výhodou této metody je jednoduchost návrhu, neboť vztahy pro impulsní odezvu filtrů typu DP (dolní propust), HP (horní propust), PP (pásmová propust) nebo PZ (pásmová zádrž), získané z IDFT, lze tabelovat. Postup návrhu lze rozepsat do několika kroků (např. pro filtry DP a HP):

1. Určení zlomové frekvence filtru f_c a její přepočet na digitální frekvenci $F_c = \frac{f_c}{f_s}$.

2. Volba délky filtru N, N musí být liché přirozené číslo.

3. Výpočet impulsní odezvy h(k) jen pro

$$-\frac{N-1}{2} \le k \le +\frac{N-1}{2} \ . \tag{4.9}$$

Tím dojde zároveň k omezení původně nekonečné impulsní odezvy na konečný počet vzorků. Impulsní odezva pro filtr typu DP má tvar

$$h(k) = 2 \cdot F_c \cdot \mathrm{sinc}\left(2 \cdot k \cdot F_c\right) \ , \tag{4.10}$$

pro filtr typu HP je rovna

$$h(k) = (-1)^k \cdot 2 \cdot F_c \cdot \mathrm{sinc}[2 \cdot k \cdot (0{,}5 - F_c)] \,, \tag{4.11}$$

kde

$$\begin{aligned} \mathrm{sinc}(x) &= \frac{\sin(\pi \cdot x)}{\pi \cdot x} \qquad \text{pro} \qquad x \neq 0 \,, \\ \mathrm{sinc}(x) &= 1 \qquad\qquad\quad \text{pro} \qquad x = 0 \,. \end{aligned} \tag{4.12}$$

4. Posun vypočtených vzorků impulsní odezvy h(k) o $\frac{N-1}{2}$ kroků vzad, tedy

$$h(k) = h\left(k - \frac{N-1}{2}\right) \qquad \text{pro} \qquad 0 \leq k \leq N-1 \,. \tag{4.13}$$

Tato operace zabezpečuje kausalitu navrhovaného filtru.

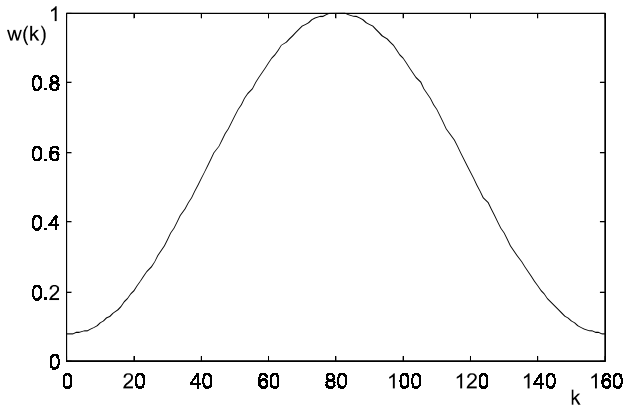
5. Aplikace okénkovací funkce, která zajistí pozvolné oříznutí impulsní odezvy h(k), a tím zabrání zkreslení spektra filtru:

$$h(k) = h(k) \cdot w(k) \qquad \text{pro} \qquad 0 \leq k \leq N-1 \,, \tag{4.14}$$

kde w(k) je zvolená okénkovací funkce, např. Hammingovo okénko (viz. obrázek 4.2), které je dáno vztahem

$$w(k) = 0{,}54 + 0{,}46 \cdot \cos\left[\frac{(N - 2 \cdot k) \cdot \pi}{N}\right] \,, \tag{4.15}$$

kde N je délka okénka.



Obr. 4.2. Hammingovo okénko délky 160 vzorků

5 Teorie automatického rozpoznávání řeči

Základní úlohou každého rozpoznávače je zařazování neznámých objektů na základě jejich popisu, tzv. obrazů, do jedné ze tříd. Tomuto procesu se také říká klasifikace. Každý nově navržený rozpoznávač musí před uvedením do provozu projít fází trénování a testování. Trénováním se rozumí předkládání obrazů objektů se známým zařazením do předem definovaných tříd, rozpoznávač si v této fázi vytvoří reprezentaci každé třídy. Při testování natrénovaného rozpoznávače se následně měří úspěšnost klasifikace obrazů objektů se známým zařazením. Soubor trénovacích obrazů je nazýván trénovací množinou, obrazy použité pro testování testovací množinou. Je žádoucí, aby obě množiny byly pokud možno co nejobsáhlejší. Výsledky testování mají určitou vypovídací hodnotu o kvalitě rozpoznávače pouze tehdy, pokud průnik trénovací a testovací množiny je prázdný, tzn. žádný obraz nebyl použit zároveň pro trénování i testování. Obecně může rozpoznávač mimo klasifikátoru obsahovat ještě další funkční části, např. pro získávání obrazů z rozpoznávaných objektů.

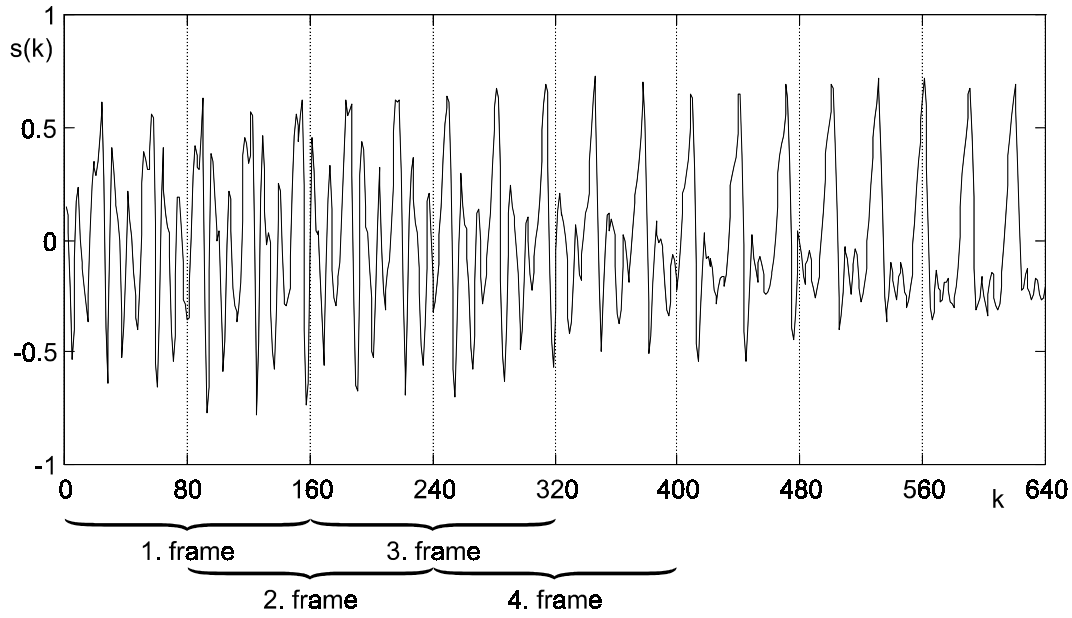
Výše uvedený popis rozpoznávače je platný i pro úlohy rozpoznávání lidské řeči. Základní klasifikovanou jednotkou lidské řeči přitom mohou být slova (většinou v rozpoznávačích izolovaných slov), nebo nižší stavební elementy řeči, jako jsou slabiky nebo hlásky (v rozpoznávačích souvislé řeči). V prvém případě je tedy definováno tolik tříd, kolik různých slov má být rozpoznáváno.

5.1 Parametrizace řečového signálu

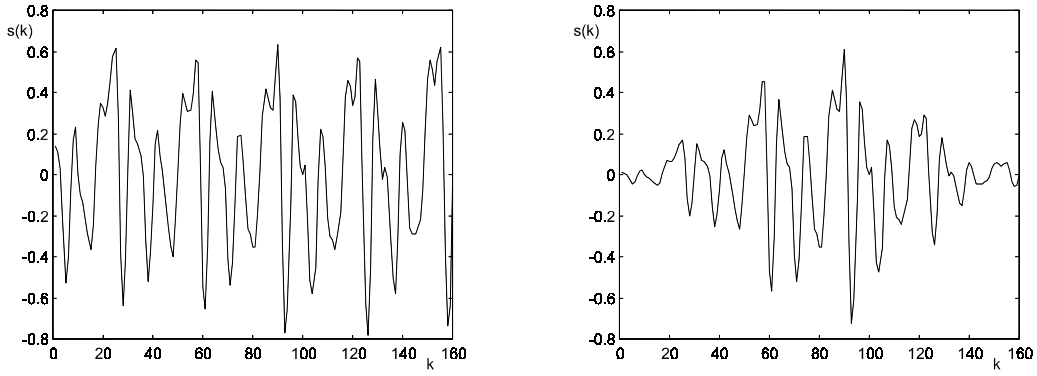
Parametrizace řečového signálu slouží k získání takového popisu signálu, který je vhodný pro jeho rozpoznávání [1], [2]. Samotné parametrizaci ještě předchází digitalizace (viz. kapitola 4.1) a segmentace signálu. I když posloupnost diskrétních vzorků, která je výsledkem digitalizace, je vlastně popisem signálu, tento popis je pro potřeby rozpoznávání zcela nevhodný, neboť obsahuje velké množství nadbytečné informace. To je také důvod, proč se po digitalizaci signál ještě segmentuje a parametrizuje vhodnými příznaky.

Zkoumáním lidské řeči bylo zjištěno, že řečový signál je po dobu několika jednotek až desítek milisekund téměř konstantní. Tento jev je dán konečnou rychlostí pohybu řečových orgánů člověka. Segmentací lze tedy signál rozdělit na krátké úseky, tzv. framy – mikrosegmenty, a jejich parametry prohlásit za konstantní. Pro každý frame potom stačí určit jen jednu hodnotu každého příznaku a díky tomu výrazně snížit objemnost potřebných dat. Často se délka framu volí 20 ms a používá se tzv. překryvu, kdy sousední framy na sebe

nenavazují, ale polovinou se překrývají (viz. obrázek 5.1). Každý frame je po vyříznutí z původního signálu ještě vážen vhodnou okénkovací funkcí, aby se zabránilo zkreslení jeho spektra (viz. obrázek 5.2). Často se pro tento účel používá Hammingovo okénko definované vztahem (4.15).



Obr. 5.1. Ukázka segmentace řečového signálu s(k) na framy délky 160 vzorků s překryvem



Obr. 5.2. Frame řečového signálu s(k) před a po vážení Hammingovým okénkem

Hlavním úkolem parametrizace řečového signálu je popsat signál tak, aby se co možná nejvíce zohlednily ty jeho části, které jsou svázány s obsahem řeči, a zároveň se co nejméně projevila osobnost konkrétního mluvčího. K tomu rozhodně nestačí jediný příznak, proto je každý frame signálu popsán celým příznakovým vektorem, který může obsahovat až čtyřicet složek, obvyklý počet je zhruba dvacet příznaků. V moderních rozpoznávačích hrají největší

roli tzv. kepstrální příznaky, které jsou často doplněny příznaky určenými v časové oblasti. Mimo těchto, tzv. statických příznaků, se používají ještě příznaky dynamické, které do popisu signálu vnášejí také informaci o vývoji statických příznaků v čase. Dynamické příznaky prvního řádu se nazývají delta příznaky, příznaky druhého řádu (vyjadřující změnu delta příznaků) delta delta příznaky. Statické příznaky jsou často ještě upravovány metodou odečítání příznakového průměru FMS (Feature Mean Subtraction), která do určité míry stírá rozdíly mezi různou kvalitou řečového signálu. Ten se může lišit právě odlišnou střední hodnotou příznaků.

5.2 Příznaky určené v časové oblasti

Snad nejpoužívanějším příznakem z této skupiny je krátkodobá energie signálu, která je pro jeden frame definována jako

$$E = \sum_{k=0}^{N-1} s(k)^2 \quad , \tag{5.1}$$

kde s(k) je vzorek parametrizovaného signálu a N je délka framu. Důležitost tohoto příznaku umocňuje fakt, že se často používá (v logaritmické podobě) k oddělování jednotlivých slov v promluvě, navíc dynamické příznaky energie jsou z hlediska rozpoznávání řeči jedny z nejdůležitějších.

Dalším možným popisem signálu je tzv. krátkodobá funkce počtu průchodů signálu nulou, která je daná předpisem

$$Z = \frac{1}{2} \cdot \sum_{k=0}^{N-2} \left| \operatorname{sgn}\left[s(k+1)\right] - \operatorname{sgn}\left[s(k)\right] \right| \ , \tag{5.2}$$

kde s(k+1) a s(k) jsou vzorky parametrizovaného signálu, N je délka framu a sgn je funkce signum definovaná jako

$$\begin{aligned} \operatorname{sgn}(x) &= 1 \quad \text{pro} \quad x \geq 0 \ , \\ \operatorname{sgn}(x) &= -1 \quad \text{pro} \quad x < 0 \ . \end{aligned} \tag{5.3}$$

Příznakem, který je vhodný např. pro určení periodicity signálu, je krátkodobá autokorelační funkce, pro její i-tý koeficient platí

$$R(i) = \sum_{k=0}^{N-1-i} s(k) \cdot s(k+i) \ , \tag{5.4}$$

kde s(k) a s(k+i) jsou vzorky parametrizovaného signálu a N je délka framu.

5.3 Kepstrální příznaky

Jak již bylo uvedeno, příznaky popisující rozpoznávaný signál by měly co nejvíce brát v úvahu ty jeho části, které nesou informaci o obsahu řeči, a naopak co nejméně části závislé na osobnosti mluvčího. Modelováním produkce lidské řeči bylo zjištěno, že na řečový signál lze zjednodušeně nahlížet jako na signál získaný konvolucí (viz. kapitola 4.2) budícího signálu a impulsní odezvy hlasového traktu. Protože přenos hlasového traktu se mění spolu s obsahem řeči a buzení je (hlavně díky hlasivkám) nejvíce ovlivněno osobou mluvčího, je účelné tyto složky oddělit. Toto právě umožňují kepstrální příznaky, neboť vycházejí z tzv. homomorfního zpracování řeči, díky kterému je možné z výsledného signálu vzniklého konvolucí vydělit jeho jednotlivé složky ve formě součtu jejich kepster. Kepstrum c(k) signálu s(k) je dáno jako zpětná diskrétní Fourierova transformace (IDFT) logaritmu absolutní hodnoty spektra signálu S(n), vztah lze zapsat jako

$$c(k) = \mathrm{IDFT}\left[\log\left|S(n)\right|\right] = \mathrm{IDFT}\left[\log\left|\mathrm{DFT}\left[s(k)\right]\right|\right] . \tag{5.5}$$

Tento předpis lze pro výpočet krátkodobého kepstra jednoho framu signálu přepsat do konkrétnější podoby

$$c(k) = \frac{1}{N}\cdot\sum_{n=0}^{N-1}\log\left|S(n)\right|\cdot e^{j\cdot\frac{2\cdot\pi\cdot n\cdot k}{N}} \qquad \text{pro} \qquad 0 \le k \le N-1 , \tag{5.6}$$

kde S(n) jsou frekvenční vzorky signálu framu určené z posloupnosti s(k) pomocí diskrétní Fourierovy transformace (DFT) a N je délka framu. Protože signál s(k) je dán konvolucí budícího signálu a impulsní odezvy hlasového traktu, vypočtené kepstrum je vlastně součtem kepster těchto dvou složek. Ze znalosti spekter buzení a impulsní odezvy lze určit, že kepstrální koeficienty s nižším indexem popisují parametry hlasového traktu, zatímco parametry budícího signálu jsou reprezentovány koeficienty s vyšším indexem. Kepstrální příznaky vypočítané přímo podle vztahu (5.6) se nazývají RCC (Real Cepstral Coefficients), často se však k výpočtu používají jiné postupy, především kvůli nižší výpočetní náročnosti. Metoda odečítání příznakového průměru FMS je při aplikaci na statické kepstrální příznaky často označována jako metoda odečítání kepstrálního průměru CMS (Cepstral Mean Subtraction).

5.3.1 Kepstrální příznaky LPCC (Linear Predictive Cepstral Coefficients)

Jako LPCC jsou označovány kepstrální příznaky určené pomocí lineární prediktivní analýzy LPC.

Za předpokladu, že produkci lidské řeči lze modelovat pomocí generátoru buzení a modelu hlasového traktu, lze řečový signál popsat vztahem

$$s(k) = -\sum_{i=1}^{Q} a_i \cdot s(k-i) + G \cdot u(k) \ , \tag{5.7}$$

kde u(k) je budící signál a G je koeficient zesílení. Tento vztah je vlastně diferenční rovnicí modelu hlasového traktu řádu Q, který je popsatelný přenosovou funkcí

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{i=1}^{Q} a_i \cdot z^{-i}} \ , \tag{5.8}$$

kde S(z) je z-obraz výstupu, U(z) je z-obraz vstupu a z^{-1} je operátor jednotkového zpoždění. Ze struktury přenosu je zřejmé, že model hlasového traktu je speciálním případem filtru typu IIR (viz. kapitola 4.2), kde čitatel je redukován pouze na nultý koeficient $b_0 = G$. Protože se tento model mění s obsahem řeči, na koeficienty přenosu lze nahlížet jako na konstantní pouze v krátkém časovém intervalu, který řádově odpovídá délce jednoho framu. Metodou LPC lze koeficienty tohoto modelu velmi přesně odhadnout, z nich lze následně určit hledané kepstrální příznaky. Zatímco kepstrum určené podle vztahu (5.6) je získáno přímo ze spektra parametrizovaného signálu, tato metoda vychází z časové oblasti a pracuje pouze se spektrální obálkou skutečného spektra. To je také důvod, proč obě metody nevedou ke stejným číselným výsledkům.

Na odhad koeficientů modelu hlasového traktu lze aplikovat metodu nejmenších čtverců, která vede na soustavu rovnic

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(Q-1) \\ R(1) & R(0) & R(1) & \cdots & R(Q-2) \\ \vdots & & & & \\ R(Q-1) & R(Q-2) & R(Q-3) & \cdots & R(0) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_Q \end{bmatrix} = \begin{bmatrix} -R(1) \\ -R(2) \\ \vdots \\ -R(Q) \end{bmatrix} , \tag{5.9}$$

kde Q je řád modelu hlasového traktu a R(0) až R(Q) jsou koeficienty krátkodobé autokorelační funkce zkoumaného framu, pro které platí vztah (5.4).

Efektivní výpočet koeficientů a_i lze provést rekurzí pomocí iterativního Durbinova algoritmu pro $1 \le i \le Q$:

$$\begin{aligned} E^{(0)} &= R(0) \ , \\ k_i &= -\frac{R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} \cdot R(i-j)}{E^{(i-1)}} \ , \\ a_i^{(i)} &= k_i \ , \\ a_j^{(i)} &= a_j^{(i-1)} + k_i \cdot a_{i-j}^{(i-1)} \qquad \text{pro} \qquad 1 \le j \le i-1 \ , \\ E^{(i)} &= \left(1 - k_i^2\right) \cdot E^{(i-1)} \ , \end{aligned} \tag{5.10}$$

kde $a_j^{(i)}$ je j-tý koeficient přenosu odhadnutý v i-tém kroku, k_i je i-tý koeficient odrazu a $E^{(i)}$ je chyba predikce. Hledané kepstrální příznaky lze z odhadnutých koeficientů a_i určit podle vztahů

$$\begin{aligned} c(k) &= -a_k && \text{pro} \quad k = 1 \, , \\ c(k) &= -a_k - \sum_{i=1}^{k-1} \left(\frac{i}{k} \right) \cdot c(i) \cdot a_{k-i} && \text{pro} \quad 2 \le k \le Q \, , \\ c(k) &= -\sum_{i=1}^{Q} \left(\frac{k-i}{k} \right) \cdot c(k-i) \cdot a_i && \text{pro} \quad k > Q \, . \end{aligned} \tag{5.11}$$

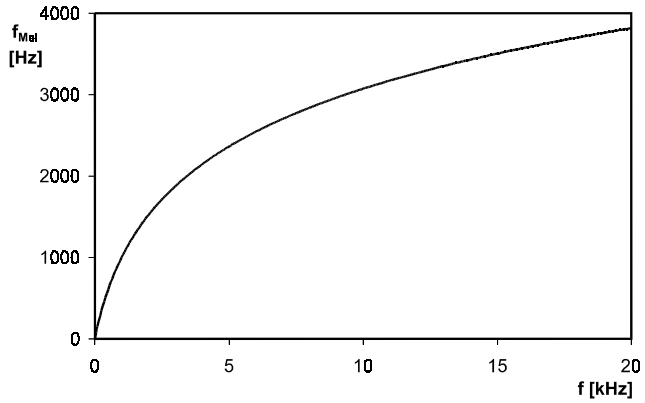
Výsledky rozpoznávání jsou při použití kepstrálních příznaků určených pomocí LPC značně závislé na vhodné volbě řádu Q modelu hlasového traktu. Běžně se používají modely 7. až 15. řádu, přičemž se nedoporučuje pracovat s kepstrálními koeficienty s indexem vyšším než je řád modelu.

5.3.2 Kepstrální příznaky MFCC (Mel-frequency Cepstral Coefficients)

Metoda pro určení kepstrálních příznaků MFCC pracuje ve frekvenční oblasti, je tedy oproti metodě LPC mnohem blíže postupu výpočtu příznaků podle vzorce (5.6). Obě frekvenční metody odlišují dva základní rozdíly. Při výpočtu MFCC se zohledňují poznatky z psychoakustiky, podle kterých lidské ucho vnímá frekvence v téměř logaritmické stupnici. Druhým rozdílem je použití diskrétní kosinové transformace (DCT) namísto zpětné diskrétní Fourierovy transformace (IDFT). Již zmíněná logaritmická stupnice se nazývá Melovská. Převod mezi lineární a Mel-frekvencí (viz. obrázek 5.3) popisuje vztah

$$f_{Mel} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \ , \tag{5.12}$$

kde f je hodnota lineární frekvence a f_{Mel} hodnota Mel-frekvence.



Obr. 5.3. Převod mezi lineární a Mel-frekvencí

Postup výpočtu MFCC příznaků je takový, že se nejprve určí spektrální vzorky framu S(n) z posloupnosti s(k) pomocí rychlé Fourierovy transformace (FFT). Následně se vytvoří banka N trojúhelníkových filtrů, která má v Melovské stupnici konstantní rozestupy Δ_m mezi centrálními frekvencemi b_m jednotlivých filtrů, a všechny filtry také mají stejnou šířku pásma. Z pohledu lineární stupnice to znamená, že rozestupy Δ_m a šířky pásem jsou konstantní pouze do frekvence zhruba 1 kHz, potom rostou v souladu se vztahem (5.12). Hodnotu Δ_m je třeba volit tak, aby se do frekvence 1 kHz vměstnalo 10 filtrů s konstantními rozestupy Δ_m mezi centrálními frekvencemi b_m. Pro rozestupy Δ_m nad frekvencí 1 kHz platí vztah

$$\Delta_m = 1{,}2 \cdot \Delta_{m-1} \ . \tag{5.13}$$

Centrální frekvence b_m nad frekvencí 1 kHz lze zcela logicky určit podle předpisu

$$b_m = b_{m-1} + \Delta_m \ . \tag{5.14}$$

Takto navržená banka filtrů se použije k výpočtu energie signálu v každém z N pásem. Výpočet je dán vztahem

$$E(m) = \sum_{n=b_m-\Delta_m}^{b_m+\Delta_m} S(n) \cdot U_{\Delta m}(n - b_m) \qquad \text{pro} \qquad 1 \le m \le N \ , \tag{5.15}$$

kde $U_{\Delta m}$ je okénko trojúhelníkového filtru, pro které platí

$$\begin{aligned} U_{\Delta m}(n) &= 1 - \frac{|n|}{\Delta_m} \qquad \text{pro} \qquad |n| < \Delta_m \, , \\ U_{\Delta m}(n) &= 0 \qquad\qquad\ \ \text{pro} \qquad |n| \ge \Delta_m \, . \end{aligned} \tag{5.16}$$

MFCC kepstrální koeficienty se nakonec získají z energií jednotlivých pásem signálu pomocí diskrétní kosinové transformace:

$$c(k) = \sqrt{\frac{2}{N}} \cdot \sum_{m=1}^{N} \log\left(E(m)\right) \cdot \cos\left(\frac{\pi \cdot k}{N} \cdot (m - 0{,}5)\right) \quad \text{pro} \quad 0 \le k \le N - 1 \ . \tag{5.17}$$

5.4 Rozpoznávače řeči

Všechny druhy rozpoznávačů řeči mají v podstatě shodnou základní strukturu, kterou lze uvažovat v podobě řetězce funkčních bloků [1], [2]. Princip jejich funkce lze jednoduše vysvětlit na rozpoznávači izolovaných slov. Prvním článkem každého rozpoznávače je vstupní zařízení, na jehož výstupu se objevuje elektrický signál reprezentující signál akustický, takovým zařízením může být například mikrofon nebo telefonní linka. Elektrický signál dále putuje do bloku, v němž se provádí digitalizace signálu (viz. kapitola 4.1), tento proces u rozpoznávačů realizovaných na počítačích zabezpečuje zvuková (nebo telefonní) karta. Digitalizovaný signál je v dalším bloku parametrizován a segmentován na jednotlivá slova (viz. kapitola 5.1). Výsledkem této operace je reprezentace každého slova vektorem framů, přičemž každý frame je popsán vektorem příznaků. Data v této podobě, nazývaná obrazem slova, potom vstupují do hlavního bloku rozpoznávače, do klasifikátoru. Právě typem klasifikátoru se jednotlivé druhy rozpoznávačů liší. Pro každé slovo, které má být rozpoznáváno, musí klasifikátor definovat jednu třídu. Rozdíly v klasifikaci jsou především dány tím, jakou reprezentaci tříd konkrétní rozpoznávač používá.

5.4.1 Klasifikace slov pomocí referencí, algoritmus DTW

Tato metoda používá k rozpoznávání slov třídy, které jsou reprezentovány pomocí tzv. referencí. Referencí se přitom rozumí obraz jednoho slova získaný parametrizací při trénování. Třída každého slova není popsána referencí jedinou, počet referencí konkrétní třídy závisí na počtu realizací příslušného slova v trénovací množině. Samotná klasifikace neznámého slova probíhá tak, že jeho obraz je porovnáván se všemi referencemi všech tříd, přitom je vyhodnocována vzdálenost těchto obrazů. Slovo je nakonec zařazeno do té třídy, pro kterou je sledovaná vzdálenost minimální.

Nutnou podmínkou této metody je stejná délka porovnávaných slov, tedy shodný počet framů jejich popisu. Potom lze vzdálenost dvou slov X a Y, resp. jejich obrazů, vyjádřit vztahem

$$D(X,Y) = \sum_{i=1}^{L} d(x_i, y_i) \ , \tag{5.18}$$

kde L je shodný počet framů slov X a Y a $d(x_i, y_i)$ je vzdálenost mezi i-tými framy slov, označovaná jako lokální. Při použití běžné Euklidovské vzdálenosti ji lze vypočítat jako

$$d(x_i, y_i) = \sqrt{\sum_{p=1}^{P} (x_{ip} - y_{ip})^2} \ , \tag{5.19}$$

kde P je počet příznaků jednoho framu a x_{ip}, resp. y_{ip}, je hodnota p-tého příznaku i-tého framu slova X, resp. Y.

V praxi však nikdy nelze dosáhnut dodržení podmínky stejné délky slov, neboť každý člověk mluví jinou rychlostí, ta se navíc může i u jednoho řečníka měnit podle situace. Řešením tohoto jevu je změna časové osy u jednoho ze slov pomocí lineární časové transformace, která umožňuje některé framy vynechat v případě zkracování slova, resp. zopakovat při jeho prodlužování. Algoritmus pracující s touto transformací je označován jako LTW (Linear Time Warping). Pro snadné porovnávání vzdáleností u různých referencí je vhodné vždy přizpůsobovat délku reference délce klasifikovaného slova, funkce w(i) zajišťující tuto lineární transformaci má potom tvar

$$w(i) = \mathrm{int}\left[\frac{J-1}{I-1} \cdot (i-1) + 1{,}5 \right] \ , \tag{5.20}$$

kde I je počet framů neznámého slova, J je počet framů reference a funkce int je definována jako funkce, která vrací celou část svého argumentu. Z předpisu funkce w(i) je zřejmé, že splňuje okrajové podmínky, které lze vyjádřit vztahy

$$\begin{aligned} w(1) &= 1 \ , \\ w(I) &= J \ . \end{aligned} \tag{5.21}$$

Výpočet vzdálenosti dvou slov X a Y lze potom provést podle vzorce velmi podobného vzorci (5.18), tedy

$$D(X,Y) = \sum_{i=1}^{I} d(x_i, y_{w(i)}) \ . \tag{5.22}$$

Podrobným zkoumáním řečového signálu byl určen ještě jeden jev, který může negativním způsobem ovlivnit kvalitu rozpoznávání. Různé realizace téhož slova totiž nemusí mít pokaždé jen jinou délku, ale také různý rytmus, neboli jiný poměr nižších stavebních

elementů řeči uvnitř slova. Tyto rozdíly lze odstranit takovou časovou transformací, která je podle potřeby schopna prodlužovat a zkracovat jednotlivé elementy slova. Toho však nelze pomocí LTW dosáhnout, proto je nutné použít takový algoritmus, který využívá nelineární časovou transformaci, algoritmus DTW (Dynamic Time Warping). Hlavní myšlenkou této metody je nalezení nejoptimálnější transformační cesty ze všech přípustných cest w, která minimalizuje vzdálenost dvou slov X a Y, pro kterou lze tedy psát

$$D(X,Y) = \min_{w}\left[\sum_{i=1}^{I} d(x_i, y_{w(i)})\right] . \tag{5.23}$$

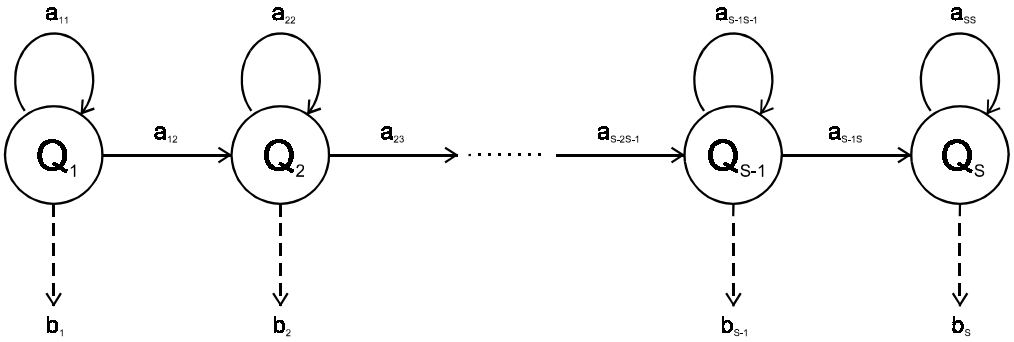
Přípustné cesty jsou ze všech možných vyděleny pomocí podmínek několika typů. Jedná se o podmínky okrajové, naprosto shodné s podmínkami u metody LTW, viz. vztahy (5.21), podmínky lokální a podmínky globální, které vymezují oblast pohybu funkce w a které lze z lokálních odvodit. Lokální podmínky se zavádí ve snaze zabránit nadměrným zásahům do časové osy obrazu reference. Docilují toho tím, že kladou požadavky na monotónnost a spojitost funkce w a zároveň předepisují určitá omezení strmosti této funkce, proto jsou také často označovány jako podmínky spojitosti a strmosti. Konkrétních typů lokálních podmínek je celá řada, navzájem se od sebe odlišují právě různými omezeními strmosti.

Hledat nejoptimálnější transformační cestu w vyčíslováním a porovnáváním vzdáleností pro každou přípustnou cestu by bylo značně neefektivní, s výhodou však lze použít Bellmanova principu optimality [6]. Tento princip zjednodušeně říká, že celkové řešení úlohy je optimální právě tehdy, pokud jsou optimální řešení všech jeho dílčích podúloh. Díky tomu je možné na výpočet aplikovat dynamické programování a celkovou vzdálenost dvou slov určovat po částech pomocí rekurze s využitím tzv. akumulované vzdálenosti, která je na konci výpočtu rovna právě vzdálenosti celkové [2].

5.4.2 Klasifikace slov pomocí skrytých markovských modelů

Rozpoznávače obsahující klasifikátor tohoto typu používají k reprezentaci tříd jednotlivých slov tzv. skryté markovské modely, často označované jako HMM (Hidden Markov Model) [2], [7]. Oproti metodě pracující s referencemi má použití modelů několik výhod. Hlavní předností této metody je reprezentace každé třídy vždy jen jedním modelem, na rozdíl od metody předchozí, u které je každé slovo reprezentováno tolika referencemi, kolik realizací slova bylo použito při trénování. V důsledku toho u referenční metody se vzrůstající robustností klasifikátoru, která souvisí právě s velikostí trénovací množiny, narůstá nejen počet referencí, ale také čas potřebný k rozpoznávání. Navíc HMM model každé třídy je

svým způsobem zobecněn, neboť v sobě částečně zahrnuje vlastnosti všech mluvčích z trénovací množiny, tím se stává univerzálnějším a použitelnějším i pro rozpoznávání řečníků, kteří nebyli zahrnuti do procesu trénování rozpoznávače. To je další výhoda oproti referenční metodě, která je více svázána s konkrétními mluvčími, neboť žádné zobecněné reprezentace nepoužívá. Naproti tomu nevýhodou klasifikace slov metodou skrytých markovských modelů je výpočetně mnohem náročnější a složitější trénování, neboť je během něj nutné určit parametry každého modelu tak, aby co nejlépe reprezentoval příslušné slovo.



Obr. 5.4. Struktura levo-pravého markovského modelu s S stavy

Struktura markovského modelu (viz. obrázek 5.4) je tvořena určitým počtem jeho stavů, přičemž možný je přechod jen mezi stavy sousedními. V oblasti rozpoznávání lidské řeči se pro všechna slova používají modely o stejném počtu stavů, téměř výhradně se jedná o tzv. levo-pravé modely, u kterých je přechod umožněn jen ze stavu s nižším indexem do stavu s indexem vyšším, tedy zleva doprava. Volba počtu stavů modelu se historicky vyvíjela, dříve se pracovalo s modely, které měly počet stavů blízký počtu framů slova. Dnes se tento počet volí mnohem menší, řádově kolem deseti stavů, neboť bylo zjištěno, že parametry několika sousedních framů se od sebe příliš neliší (viz. kapitola 5.1) a lze je tedy bez velké ztráty informace zastoupit jen jedním stavem. Každý stav modelu (viz. obrázek 5.5) je potom charakterizován několika parametry, které se určí ve fázi trénování. Jeden z parametrů stavu určuje pravděpodobnost setrvání modelu v tomto stavu, druhý pravděpodobnost přechodu modelu do stavu vyššího, tyto parametry se označují jako a_{ss} a a_{ss+1} a jsou definovány jako

$$a_{ss+1} = \frac{K}{N_s} \ , \qquad a_{ss} = 1 - a_{ss+1} \ , \tag{5.24}$$

kde K je počet trénovacích realizací slova, kterému přísluší model se stavem s, a N_s je počet framů reprezentovaných tímto stavem. Dalšími důležitými parametry jsou střední hodnoty

a rozptyly jednotlivých příznaků framů, které jsou daným stavem reprezentovány, protože mu byly během trénování přiřazeny. Informace o středních hodnotách příznaků framů stavu s je uložena ve vektoru $\overline{x}_s$

$$\overline{x}_s = \begin{bmatrix} \overline{x}_{s1} \\ \overline{x}_{s2} \\ \vdots \\ \overline{x}_{si} \\ \vdots \\ \overline{x}_{sP} \end{bmatrix}, \tag{5.25}$$

kde P je počet příznaků každého framu a $\overline{x}_{si}$ je střední hodnota i-tého příznaku framů stavu s, která je definována jako

$$\overline{x}_{si} = \frac{1}{N_s} \cdot \sum_{n=1}^{N_s} x_{ni} \ , \tag{5.26}$$

kde N_s je počet framů reprezentovaných stavem s a x_{ni} je hodnota i-tého příznaku n-tého framu stavu s. Hodnoty rozptylů jednotlivých příznaků framů stavu s jsou uloženy na diagonále tzv. kovariační matice Σ_s, kterou lze uvažovat ve tvaru

$$\Sigma_s = \begin{bmatrix} \sigma_{s1}^{\,2} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_{s2}^{\,2} & & & & \\ \vdots & & \ddots & & & \\ 0 & & & \sigma_{si}^{\,2} & & \\ \vdots & & & & \ddots & \\ 0 & & & & & \sigma_{sP}^{\,2} \end{bmatrix}, \tag{5.27}$$

kde P je opět počet příznaků framu a $\sigma_{si}^{\,2}$ je rozptyl i-tého příznaku framů stavu s, který je definován jako

$$\sigma_{si}^{\,2} = \frac{1}{N_s} \cdot \sum_{n=1}^{N_s} \left(x_{ni} - \overline{x}_{si} \right)^2 \ , \tag{5.28}$$

kde N_s je opět počet framů reprezentovaných stavem s, x_{ni} je hodnota i-tého příznaku n-tého framu stavu s a $\overline{x}_{si}$ je střední hodnota i-tého příznaku framů stavu s. Se znalostí parametrů $\overline{x}_s$ a Σ_s lze pro každý stav modelu definovat tzv. pravděpodobností výstupní funkci, pomocí které lze vyčíslit hodnotu pravděpodobnosti, že určitý frame přísluší tomuto stavu.

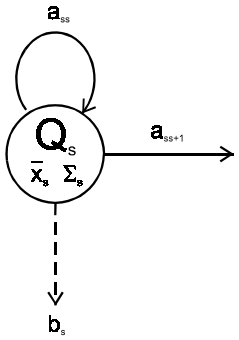
Zmíněná funkce je označována jako $b_s(x)$ a je definována vztahem

$$b_s(x) = \frac{1}{\sqrt{(2\cdot\pi)^P \cdot \det \Sigma_s}} \cdot e^{\left[-\frac{1}{2}\cdot(x-\bar{x}_s)^T \cdot \Sigma_s^{-1} \cdot (x-\bar{x}_s)\right]}, \tag{5.29}$$

kde x je vektor příznaků framu, u kterého se testuje příslušnost ke stavu s. Některé klasifikátory pracují se složitějšími, tzv. vícemixturovými modely, u kterých nejsou stavy popsány jen jedním vektorem $\bar{x}_s$ a maticí Σ_s. Každý stav takového modelu je reprezentován více složkami, tzv. mixturami, parametry $\bar{x}_s$ a Σ_s jsou potom určeny zvlášť pro každou jeho složku. Předpis pro pravděpodobnostní výstupní funkci má v tomto případě tvar

$$b_s(x) = \sum_{m=1}^{M} c_{sm} \cdot \frac{1}{\sqrt{(2\cdot\pi)^P \cdot \det \Sigma_{sm}}} \cdot e^{\left[-\frac{1}{2}\cdot(x-\bar{x}_{sm})^T \cdot \Sigma_{sm}^{-1} \cdot (x-\bar{x}_{sm})\right]}, \tag{5.30}$$

kde M je počet mixtur stavu modelu a c_{sm} je váhový koeficient m-té mixtury.



Obr. 5.5. Struktura jednomixturového stavu skrytého markovského modelu

Klasifikace neznámého slova probíhá určováním míry pravděpodobnosti, že obraz tohoto slova patří k modelu reprezentujícímu nějakou třídu. Tato pravděpodobnost se postupně určuje pro modely všech tříd a slovo je nakonec zařazeno do té třídy, pro kterou je sledovaná pravděpodobnost maximální. Pro konkrétní slovo X a model M ji lze určit podle předpisu

$$P(X,M) = \max_f \left[\prod_{i=1}^{I} a_{f(i-1)\ f(i)} \cdot b_{f(i)}(x_i) \right], \tag{5.31}$$

kde pravděpodobnost přechodu modelu do počátečního stavu $a_{f(0)\ f(1)}$ je rovna jedné, I je počet framů neznámého slova a f je funkce přiřazující jednotlivým framům slova stavy modelu. Mezi všemi přípustnými funkcemi f se přitom hledá ta nejoptimálnější, která pravděpodobnost (5.31) maximalizuje. Přípustnost funkce je dostatečně určena již samotnou

strukturou skrytých markovských modelů a není tedy třeba na rozdíl od metody DTW definovat pro ni žádné další podmínky.

Analogie mezi metodou DTW a skrytými markovskými modely je zcela zřejmá, proto i zde lze při hledání nejoptimálnější přiřazující funkce f použít Bellmanova principu optimality. Na výpočet lze tedy aplikovat dynamické programování a výslednou pravděpodobnost určovat opět po částech pomocí rekurze. Tento postup je znám pod názvem Viterbiho algoritmus [2].

6 Realizace simulátoru telefonní linky

Jak již bylo řečeno v úvodu (viz. kapitola 3), mým úkolem bylo vytvořit software simulující rušivé vlivy přenosové cesty akustického signálu při rozpoznávání řeči po telefonu. Prostředků, kterými lze dosáhnout úpravy akustického signálu do podoby odpovídající přenosu po telefonní lince, je jistě celá řada. Rozhodl jsem se zvolit dva základní přístupy, které se jevily jako nejnadějnější s tím, že posléze provedu porovnání úspěšnosti obou těchto metod, případně ověřím také úspěšnost jejich kombinace.

6.1 Metoda první - banka filtrů

Tento přístup je založen na analýze zkušebních telefonních nahrávek. V nahrávkách je třeba odhalit jednotlivé rušivé vlivy a vytvořit software, který bude vstupní signál filtrovat pomocí banky filtrů, kde každý filtr bude reprezentovat některý z těchto rušivých vlivů.

6.1.1 Pořízení zkušebních telefonních nahrávek

Prvním krokem této metody je pořízení zkušebních telefonních nahrávek pomocí počítačové telefonní karty. Je důležité, aby těchto nahrávek bylo co možná nejvíce a aby byly co nejrozmanitější, tedy z různých telefonních sítí a různých telefonních přístrojů. Jedině potom se v nich mohou promítnout všechny rušivé vlivy přítomné na telefonní lince. Vhodné je také nahrávání provést na stejném hardwaru a ve stejném formátu dat, který bude následně použit v kombinaci s rozpoznávačem k ověření úspěšnosti metody.

Konkrétně jsem nahrávání prováděl pomocí telefonní karty Dialogic D-21 ve formátu: vzorkovací frekvence 8 kHz, 8 bitů na vzorek, kódování μ-law. Ihned při ukládání byla data převáděna do formátu: vzorkovací frekvence 8 kHz, 16 bitů na vzorek, kódování PCM. Takto jsem pořídil zhruba dvacet nahrávek mluveného slova o délce deseti až dvaceti sekund. Zkušební hovory byly uskutečňovány z pevných i mobilních telefonních sítí. Jako pevná síť byla konkrétně použita síť Českého Telecomu, a.s., ve které byly provedeny jak hovory místní, v rámci téhož telefonního obvodu, tak hovory dálkové, mezi dvěma telefonními obvody. Z mobilních byly použity digitální sítě standardu GSM všech tří operátorů působících v České republice, tedy síť Eurotel GSM (Eurotel Praha, spol. s r.o.), síť Oskar (Český Mobil, a.s.) a síť T-Mobile (RadioMobil, a.s.). Analogová mobilní síť systému NMT, v České republice provozovaná společností Eurotel Praha, spol. s r.o. pod názvem Eurotel T!P, do nahrávek zařazena nebyla. Ke každému telefonátu byl použit jiný telefonní přístroj.

6.1.2 Analýza zkušebních telefonních nahrávek

Dospěl jsem k závěru, že analýzu zkušebních nahrávek je vhodné provádět ve dvou fázích. V první fázi je třeba zjistit, jaké rušivé vlivy se v nahrávkách vyskytují a rozhodnout, které z nich je možné zanedbat a které naopak v druhé fázi analýzy podrobněji prozkoumat.

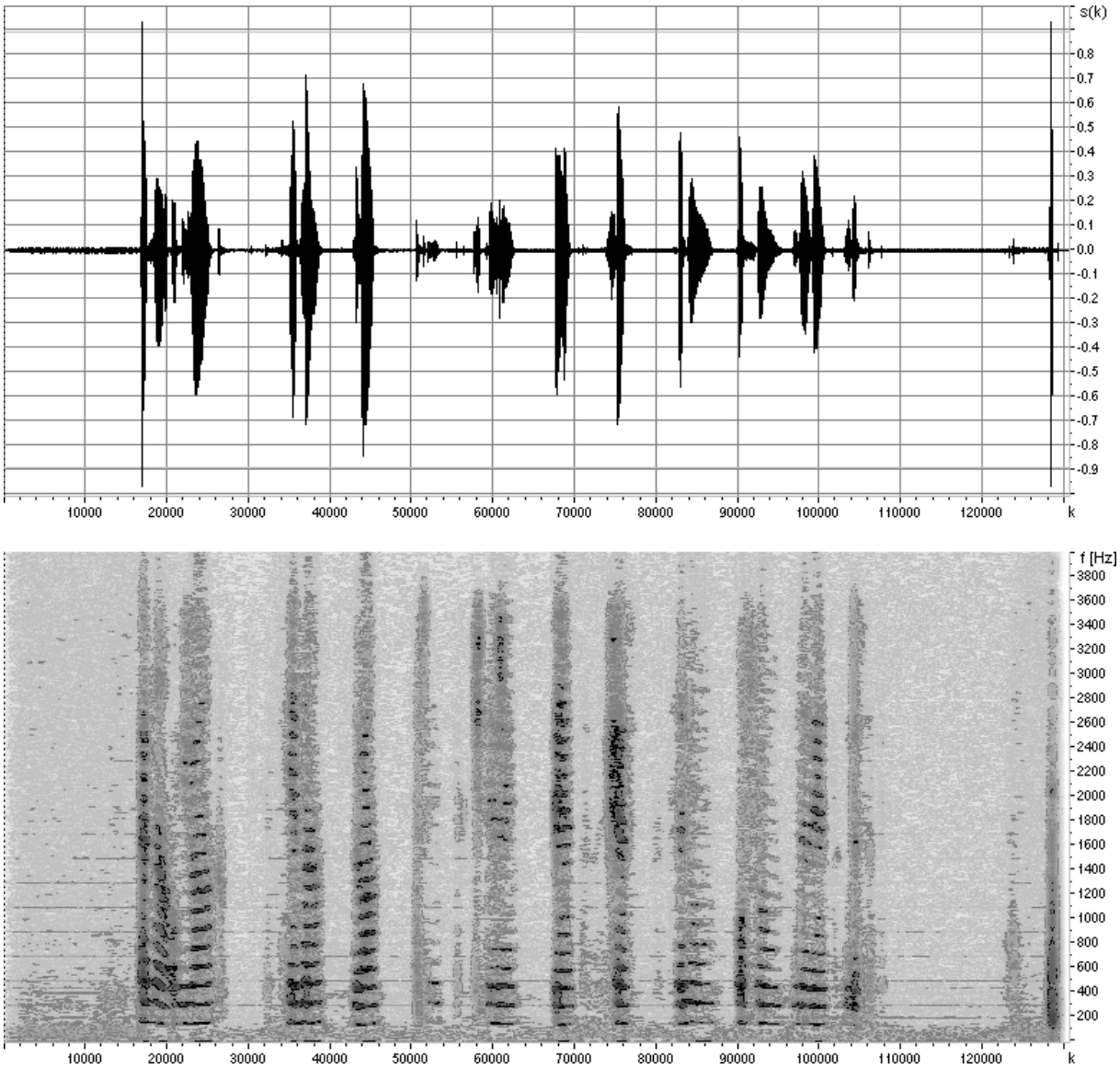
První část analýzy jsem prováděl pomocí pečlivého poslechu a pomocí spektrogramů zaznamenaných signálů. V druhé části jsem také používal převážně spektrogramy, navíc jsem pracoval i s amplitudovými frekvenčními charakteristikami. Veškeré zjištěné údaje o nahrávkách jsem shromažďoval odděleně pro pevné a mobilní telefonní sítě, neboť v této fázi ještě nebylo zřejmé, zda rozdíly mezi sítěmi budou zanedbatelné, nebo bude třeba je také zohlednit.

Z první fáze analýzy záznamů vyplynulo, že veškeré nahrávky byly frekvenčně omezené, obsahovaly šum a většina také brum, tedy rušivé vlivy, které lze u telefonní linky očekávat. Naopak ozvěna, tzv. echo, někdy také přítomná v telefonním hovoru, nebyla pozorovatelná v žádné z nahrávek. To zřejmě souvisí s důvody vzniku tohoto jevu. Pokud spolu totiž hovoří dva účastníci pomocí telefonních přístrojů, může dojít k situaci, kdy přístroj svým mikrofonem zachytí zvuk vycházející z reproduktoru a posílá jej po telefonní lince zpět. Vlivem toho potom mluvčí slyší ve sluchátku svůj vlastní hlas. Pokud je hovor uskutečněn v pevné síti, kde je zpoždění signálu minimální, tento jev nepůsobí příliš rušivě a mnohdy si ho telefonující ani neuvědomí. Horší situace nastává při použití digitálního mobilního telefonu, kde je zpoždění signálu větší, především vlivem kódování a zpětného dekódování. Mluvčímu se potom jeho hlas vrací zpět do sluchátka s již znatelným zpožděním, a tím dochází k efektu ozvěny. Vzhledem k tomu, že pořizování zkušebních nahrávek probíhalo pomocí počítačové telefonní karty, která navíc pracuje pouze v simplexním režimu, a to ve směru od mluvčího k počítači, je pochopitelné, že se tento rušivý efekt neprojevil.

V druhé fázi jsem se tedy zaměřil pouze na přesnější určení frekvenčního omezení nahrávek a na vlastnosti přítomného šumu a brumu. Hned zpočátku analýzy se ukázalo, že teoretické telefonní pásmo, v literatuře často uváděné jako 300 Hz až 3,4 kHz, neodpovídá současné realitě, neboť téměř všechny nahrávky byly frekvenčně bohatší. Je pochopitelné, že z nahrávek nelze určit frekvence vymezující telefonní pásmo naprosto přesně, lze je pouze odhadnout. Proto jsem se rozhodl hornofrekvenční i dolnofrekvenční omezení popsat každé pomocí dvou frekvencí, a tím určit hranice tzv. propustného a nepropustného pásma. Dolní i horní hranici propustného pásma jsem určil tak, že jsem ve spektrogramu hledal frekvenci, od které začíná amplituda signálu viditelně slábnout. Obě hranice nepropustného pásma jsem

určil jako frekvence, na kterých je velikost užitečného signálu již tak malá, že se tento signál začíná ztrácet v šumu. Tyto hodnoty jsem hledal v každém ze záznamů a hodnoty výsledné potom určil jako jejich aritmetické průměry. Výsledky této analýzy jsou shrnuty v tabulce 6.1.

U několika záznamů došlo k jevu, kdy v hlasitých pasážích nebyl signál shora omezen téměř vůbec, tedy výraznější hodnoty amplitudy byly viditelné i na frekvencích těsně pod hodnotou 4 kHz. Příklad takového záznamu je vidět na obrázku 6.1. Tento jev jsem při analýze zanedbal, při následném návrhu softwaru jej ale uvažoval.



Obr. 6.1. Časový průběh signálu (nahoře) a spektrogram (dole) jednoho ze záznamů

Tab. 6.1. Nalezené hranice propustného a nepropustného pásma

	Pevná telefonní síť	**Mobilní telefonní síť**
Propustné pásmo	170 Hz až 3655 Hz	200 Hz až 3585 Hz
Nepropustné pásmo	0 Hz až 140 Hz, 3735 Hz až 4000 Hz	0 Hz až 150 Hz, 3655 Hz až 4000 Hz

Z analýzy spektrogramů záznamů lze dále konstatovat, že šum se nachází v prakticky celém spektru 0 Hz až 4 kHz, brum se vyskytuje v pásmu od 0 Hz do 150 Hz (viz. obrázek 6.1). Tyto hodnoty platí pro všechny záznamy, výjimku tvoří jen ty, u kterých se brum téměř neprojevil. K určování velikosti šumu a brumu jsem použil ze záznamů jen ty jejich části, které neobsahovaly žádný užitečný signál, tedy pouze pauzy mezi slovy, neboť neexistuje žádný jednoduchý způsob pro oddělení těchto rušivých vlivů od užitečného signálu. Takto upravený signál jsem dále filtrací strmými filtry se zlomovou frekvencí 150 Hz rozdělil na dvě části. Část filtrovanou dolnofrekvenční propustí jsem použil pro zkoumání brumu, část získanou pomocí hornofrekvenční propusti ke zkoumání šumu. U obou signálů jsem potom provedl výpočet průměrné hodnoty velikosti šumu, resp. brumu, podle vztahu

$$\overline{|s|} = \frac{1}{N} \cdot \sum_{k=0}^{N-1} |s(k)| \ , \tag{6.1}$$

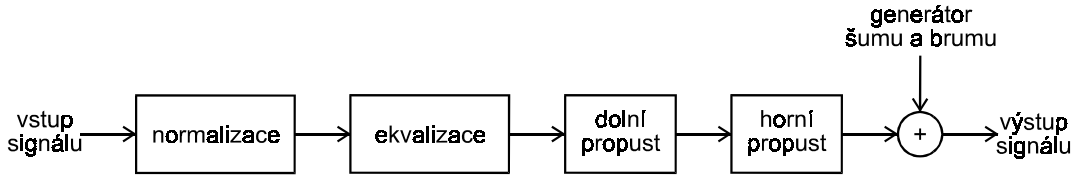
kde s(k) je hodnota k-tého vzorku signálu a N je počet těchto vzorků. Výsledné hodnoty průměrných velikostí šumu a brumu, které jsou zachyceny v tabulce 6.2, jsem opět určil jako aritmetické průměry hodnot jednotlivých záznamů. V tabulce jsou tyto velikosti uvedené v absolutním formátu, tedy pro 16-ti bitový záznam v rozsahu od 0 do 32 767. Mimo jiné je z tabulky také zřejmé, že při softwarové realizaci šumu a brumu nebude třeba rozlišovat pevné a mobilní telefonní linky, neboť nalezené hodnoty jsou u nich téměř shodné.

Tab. 6.2. Nalezené vlastnosti šumu a brumu

		Pevná telefonní síť	**Mobilní telefonní síť**
Šum	**frekvenční pásmo**	0 Hz až 4 kHz	0 Hz až 4 kHz
	průměrná velikost	58	61
Brum	**frekvenční pásmo**	0 Hz až 150 Hz	0 Hz až 150 Hz
	průměrná velikost	52	53

6.1.3 Návrh softwaru

Pro realizaci simulačního softwaru jsem zvolil prostředí Microsoft Visual C++ 6.0 a program vytvořil jako konzolovou aplikaci Win32 s možností konfigurace z příkazového řádku pomocí textového souboru. Program lze rozdělit do několika bloků, kterými postupně vstupní signál načtený ze souboru prochází (viz. obrázek 6.2).

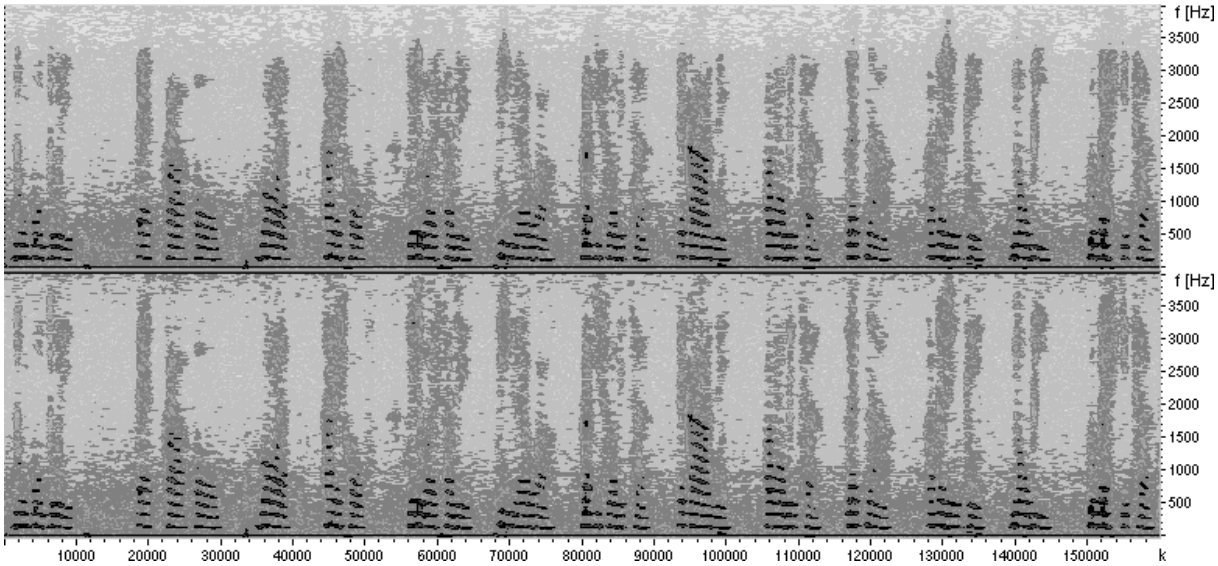


Obr. 6.2. Blokové schéma simulačního programu

První úpravou vstupního signálu je tzv. normalizace. Ta zabezpečuje změnu hlasitosti signálu tak, aby nejhlasitější místo tohoto signálu bylo rovno požadované hodnotě. Vhodnou normalizací (standardně se používá normalizace na 98 % maximální hlasitosti) tedy lze nejen odstranit rozdíly mezi různě hlasitými vstupními záznamy, ale hlavně mnohem hospodárněji využít celý kvantizační rozsah, který je dán hodnotami od $-\frac{2^n}{2}$ do $+\frac{2^n}{2}-1$, kde n je počet bitů každého vzorku (viz. kapitola 4.1).

V dalším kroku je třeba signál ekvalizovat, neboť během zkoumání pokusných záznamů pořízených pomocí různých mikrofonů a zvukových karet jsem narazil na velmi nepříjemnou skutečnost. Většina těchto záznamů (8 kHz, 16 bitů, PCM) byla shora frekvenčně omezena mnohem silněji, než záznamy pořízené telefonní kartou z telefonní linky, znatelné zeslabení amplitudy bylo viditelné již od frekvence 3,5 kHz (viz. obrázek 6.3). Tento jev lze vysvětlit tím, že zvuková karta pracující v režimu 8 kHz zařadí do cesty signálu antialiasingový filtr (viz. kapitola 4.1), který má za úkol odstranit z něj všechny složky nad frekvencí 4 kHz, zároveň však poměrně významně zeslabí i spektrum signálu od zmíněných 3,5 kHz výše. Pokud má ale program provádět úpravu signálu na telefonní pásmo, musí pracovat se signálem plnohodnotným, tedy takovým, který vyplňuje celé frekvenční pásmo od 0 do 4 kHz. Toho lze dosáhnout právě ekvalizací. Konkrétně ji navržený program provádí tím, že signál rozdělí nepříliš strmou dolní propustí se zlomovou frekvencí 3,5 kHz na dvě části. V každé z nich určí průměrnou hodnotu signálu a z poměru těchto hodnot stanoví koeficient, kterým bude část 3,5 až 4 kHz zesílena. Mnoha zkouškami jsem ověřil, že frekvenčně bohaté vstupní signály program uvedenou ekvalizací téměř vůbec nezmění,

naopak u signálů s již popsaným jevem podpoří vyšší kmitočty, a tím upraví jejich spektrum do požadované podoby.



Obr. 6.3. Spektrogram části mikrofonního záznamu před (nahoře) a po (dole) ekvalizaci

Následně je v programu provedena stěžejní úprava signálu, jeho frekvenční omezení na úroveň odpovídající telefonnímu pásmu. K tomu jsou použity dva filtry s konečnou impulzní odezvou (FIR), dolní a horní propust. Návrh těchto filtrů, tedy výpočet hodnot vzorků jejich impulsních odezev, je proveden přímo v programu pomocí tzv. metody oken (viz. kapitola 4.2). Celou filtraci by samozřejmě šlo provést jedním filtrem typu pásmová propust, úmyslně jsou však použity filtry dva, aby bylo možné nezávisle u každého z nich volit jeho strmost. Samotný výpočet filtrovaného signálu je prováděn v diskrétní časové oblasti pomocí konvoluce vzorků signálu a vzorků impulzní odezvy filtru (viz. kapitola 4.2).

Pro simulaci již zmíněného jevu, kdy v nejhlasitějších pasážích nemá dojít k téměř žádnému omezení signálu shora (viz. obrázek 6.1), je v programu použit následující postup. Před samotnou filtrací dolní propustí je v paměti programu uchován původní nefiltrovaný signál, od něj je následně výsledek filtrace odečten. Takto získaný signál je vlastně roven signálu, který dolní propust filtrací odstranila. Ten je poté zeslaben normalizací na velmi malou hodnotu, např. 2,5 %, a přičten zpět k filtrovanému signálu. Tím je vlastně provedena určitá kompenzace účinku dolní propusti.

Poslední část programu zabezpečuje zarušení upravovaného signálu šumem a brumem. Šumový signál se zvolenou průměrnou velikostí je vytvořen pomocí generátoru náhodných čísel a je k užitečnému signálu přičten. Stejným způsobem je vytvořen i brum, ten je však

před přičtením k signálu ještě filtrován vhodně zvolenou dolní propustí, neboť nemá být přítomen v celém frekvenčním pásmu, ale jen v jeho určité spodní části.

Je zřejmé, že optimální nastavení simulačního programu lze získat až z výsledků rozpoznávacích experimentů. Hrubé nastavení jednotlivých parametrů, použitelné jako výchozí pro rozpoznávací experimenty, však lze nalézt i bez těchto experimentů. Stačí k tomu pořídit několik pokusných mikrofonních nahrávek, každou nahrávku upravit navrženým simulačním softwarem a následně ji analyzovat stejnými prostředky jako zkušební telefonní nahrávky. Při vhodném nastavení parametrů by měly být výsledky analýzy telefonních a upravených mikrofonních nahrávek téměř shodné. Touto cestou jsem došel k parametrům uvedeným v tabulkách 6.3 a 6.4.

Tab. 6.3. Nalezené parametry různé pro pevnou a mobilní telefonní síť

		Pevná telefonní síť	**Mobilní telefonní síť**
Dolní propust	**zlomová frekvence**	3550 Hz	3480 Hz
	délka filtru	41 vzorků	45 vzorků
Horní propust	**zlomová frekvence**	200 Hz	225 Hz
	délka filtru	181 vzorků	151 vzorků

Tab. 6.4. Nalezené parametry shodné pro pevnou a mobilní telefonní síť

Ekvalizace	**zlomová frekvence filtru**	3500 Hz
	délka filtru	21 vzorků
Šum	**velikost**	60
Brum	**velikost**	53
	zlomová frekvence filtru	50 Hz
	délka filtru	65 vzorků

6.1.4 Popis konfigurace softwaru

Syntaxe pro spuštění navržené konzolové aplikace je následující:

```
MICtoTEL_bf.exe <vstupní_soubor> <výstupní_soubor> <konfigurační_soubor>
```

Program akceptuje vstupní zvukové soubory ve formátu: vzorkovací frekvence 8 kHz, 16 bitů na vzorek, kódování PCM, mono. Pokud je jako vstupní zvolen soubor typu WAV, program ihned po spuštění zkontroluje podle hlavičky jeho formát a pokud se neshoduje s formátem výše uvedeným, aplikace se ukončí. U souborů jiných typů program zmíněný formát předpokládá, žádnou jeho kontrolu neprovádí a s celým obsahem souboru pracuje jako se zvukovými daty. Výstupní soubor je vytvořen vždy ve shodném formátu s formátem vstupního souboru.

Konfigurační soubor je běžným textovým souborem, jehož obsah je požadován ve formátu:

```
normalizace 98
ekvalizace_zlomova_frekvence_filtru 3500
ekvalizace_delka_filtru 21
dolni_propust_zlomova_frekvence 3550
dolni_propust_delka 41
dolni_propust_kompenzace 2.5
horni_propust_zlomova_frekvence 200
horni_propust_delka 181
sum_velikost 60
brum_velikost 53
brum_zlomova_frekvence_filtru 50
brum_delka_filtru 65
```

Pokud v souboru některý z řádků chybí, není v něm uvedena žádná hodnota parametru, nebo obsahuje hodnotu nepřípustnou, příslušný efekt je vyřazen a program jej neprovádí. Podrobný přehled jednotlivých parametrů a jejich přípustných hodnot je uveden v tabulce 6.5.

Tab. 6.5. Přehled parametrů a jejich přípustných hodnot

Označení parametru	Typ	Přípustné hodnoty
normalizace	reálné číslo	> 0
ekvalizace_zlomova_frekvence_filtru	celé číslo	> 0 ≤ 4000
ekvalizace_delka_filtru	celé liché číslo	> 0
dolni_propust_zlomova_frekvence	celé číslo	> 0 ≤ 4000
dolni_propust_delka	celé liché číslo	> 0
dolni_propust_kompenzace	reálné číslo	> 0
horni_propust_zlomova_frekvence	celé číslo	> 0 ≤ 4000
horni_propust_delka	celé liché číslo	> 0
sum_velikost	celé číslo	> 0
brum_velikost	celé číslo	> 0
brum_zlomova_frekvence_filtru	celé číslo	> 0 ≤ 4000
brum_delka_filtru	celé liché číslo	> 0

6.2 Metoda druhá – identifikovaný model

Na celou přenosovou cestu telefonní linky lze nahlížet jako na jednu soustavu. Protože se na této cestě kromě užitečného signálu vyskytuje také šum, tedy náhodná veličina, nelze soustavu popsat pomocí jediné přenosové funkce. K popisu však lze s výhodou použít model uvažující působení náhodného signálu. Takový model v sobě zahrnuje dvě přenosové funkce, jednu pro popis deterministické části soustavy, druhou pro popis náhodné, tzv. stochastické, části soustavy. Právě na tomto přístupu je založena druhá zvolená metoda pro simulaci telefonní linky.

Pro identifikaci modelu, přesněji řečeno pro odhad jeho parametrů, je třeba identifikovat jeho dílčí přenosové funkce z vhodných telefonních nahrávek, tzv. testovacích signálů.

6.2.1 Pořízení telefonních nahrávek pro identifikaci

Telefonní nahrávky použitelné pro identifikaci lze pořídit pomocí dvou počítačových telefonních karet. Stejně jako v případě první metody je důležité, aby nahrávek byl co možná největší počet (s různými testovacími signály) a aby byly pořízeny z různých telefonních sítí. Platí zde také pravidlo, že je vhodné nahrávání provést na stejném hardwaru, který bude použit k rozpoznávacím experimentům.

K získání nahrávek jsem použil dvě telefonní karty, jednou z nich byla opět karta Dialogic D-21, jako druhá posloužila karta Creative Labs PhoneBlaster. Nahrávání probíhalo tak, že první kartou byly na linku pouštěny různé testovací signály, druhá karta je na opačné straně linky zaznamenávala. Konkrétně byly jako testovací použity tyto signály: několik různých druhů rozmítaných signálů, bílý šum, úryvek mluveného slova, úryvek hudby a série jednotkových skoků.

Bohužel bylo toto nahrávání poznamenáno různými technickými omezeními. Jedním z nich byla rozdílnost formátů, v kterém obě karty pracují. Nahrávky na výstupu z telefonní linky byly pořízeny ve formátu: 8 kHz, 16 bitů na vzorek, kódování PCM, oproti tomu formát vstupních nahrávek byl: 8 kHz, 8 bitů na vzorek, kódování μ-law. Z toho důvodu musely být nahrávky před samotnou identifikací převáděny na shodný formát, v tomto případě na PCM. Mnohem závažnějším technickým omezením byla možnost uskutečnit nahrávání pouze na jedné telefonní lince, navíc jen v rámci jedné pobočkové ústředny. Je logické, že vlastnosti takové linky nemohou odpovídat vlastnostem běžných linek pevné či mobilní telefonní sítě, tedy ani rušivé vlivy na těchto linkách nemusejí být shodné.

6.2.2 Odhad parametrů modelu

Číslicových modelů uvažujících působení náhodného signálu je celá řada, navzájem se od sebe liší svou strukturou, tedy typem přenosových funkcí. Před samotným odhadem parametrů je proto třeba nejprve zvolit druh modelu a také řády všech polynomů, které se v něm vyskytují. Poté je nutné vybrat vhodný testovací signál, pomocí kterého se parametry modelu identifikují. Metod pro odhad parametrů modelu existuje mnoho, zdaleka nejpoužívanější je metoda nejmenších čtverců [8], [9].

Pro popis přenosu telefonní linky jsem zvolil snad vůbec nejpoužívanější model, model nazývaný jako ARX (Auto-Regressive with eXogenous variable) (viz. obrázek 6.4) [8], [9]. Jeho struktura je taková, že přenos popisující deterministickou část modelu má tvar

$$\frac{B(z)}{A(z)} = \frac{b_0 + b_1 \cdot z^{-1} + b_2 \cdot z^{-2} + ... + b_M \cdot z^{-M}}{1 + a_1 \cdot z^{-1} + a_2 \cdot z^{-2} + ... + a_N \cdot z^{-N}} \quad , \tag{6.2}$$

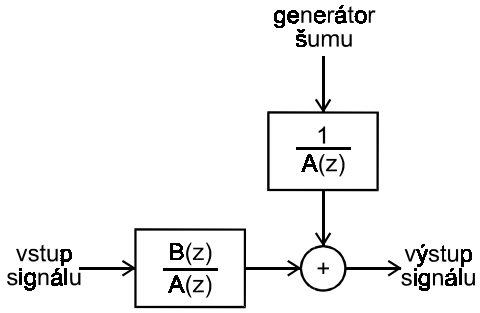
přenos aproximující účinky stochastického signálu je dán vztahem

$$\frac{1}{A(z)} = \frac{1}{1 + a_1 \cdot z^{-1} + a_2 \cdot z^{-2} + ... + a_N \cdot z^{-N}} \quad . \tag{6.3}$$

Z přenosových funkcí je zřejmé, že v obou případech se jedná o filtry typu IIR (viz. kapitola 4.2). Model ARX má dva vstupy, do filtru s přenosem $\frac{B(z)}{A(z)}$ vstupuje užitečný signál, zatímco vstupem filtru $\frac{1}{A(z)}$ je signál stochastický, bílý šum. Výstupní signál z modelu je dán součtem signálů vystupujících ze zmíněných dvou přenosů. Diferenční rovnice popisující celý model má tedy tvar

$$y(k) + ... + a_N \cdot y(k - N) = b_0 \cdot x(k) + ... + b_M \cdot x(k - M) + \nu(k), \qquad (6.4)$$

kde y(k) je výstup modelu, s(k) je deterministický vstup modelu a ν(k) je stochastický vstup modelu.

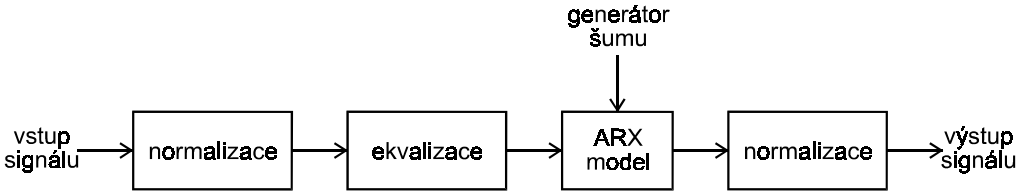


Obr. 6.4. Struktura modelu ARX

Vlastní identifikaci jsem prováděl pomocí softwaru MATLAB 5.3 (R11), který pro odhad parametrů modelu využívá již zmíněnou metodu nejmenších čtverců. Protože nelze předem bez rozpoznávacích experimentů určit, který testovací signál a jaké řády polynomů budou pro identifikaci nejvhodnější, postupně jsem použil všechny dříve vyjmenované testovací signály, s každým ze signálů jsem navíc provedl více identifikací s různými řády polynomů A(z) a B(z). Jejich hodnoty jsem nejprve volil zcela náhodně, až podle analýzy výstupů ze simulačního softwaru jsem určil intervaly hodnot, ve kterých je vhodné se pohybovat při určování řádů polynomů pro další identifikace. Tímto postupem jsem tedy obdržel mnoho variant ARX modelu telefonní linky.

6.2.3 Návrh softwaru

Stejně jako v případě první metody i zde jsem simulační software vytvořil jako konzolovou aplikaci s možností konfigurace z příkazového řádku, hodnoty koeficientů polynomů A(z) a B(z) se do programu předávají pomocí dvou textových souborů. Jednotlivé bloky programu jsou patrné z obrázku 6.5.



Obr. 6.5. Blokové schéma simulačního programu

První dvě fáze úpravy vstupního signálu se neliší od programu popsaného v kapitole 6.1.3. Signál je tedy opět nejprve normalizován na zadanou hodnotu hlasitosti a následně je u něj z již zmíněných důvodů provedena ekvalizace.

Takto upravený signál posléze vstupuje do modelu ARX přenosové cesty telefonní linky, ve kterém je frekvenčně omezen a zároveň zarušen. Mimo tohoto signálu musí do modelu ARX také vstupovat stochastický signál vytvořený pomocí generátoru náhodných čísel. Samotný výpočet signálu vystupujícího z modelu je prováděn v diskrétní časové oblasti rekurzivně pomocí tzv. Pierceova algoritmu (viz. kapitola 4.2).

Poslední úpravou signálu v tomto simulačním softwaru je jeho opětovná normalizace. Ta je potřebná z toho důvodu, neboť jak jsem v počátku zjistil analýzou výstupních signálů, identifikovaný model téměř nikdy nepopisuje věrně statické vlastnosti telefonní linky a následkem toho dochází při průchodu signálu modelem k poměrně výrazné změně jeho hlasitosti, přesněji řečeno k jeho zeslabení. Tento nepříznivý jev normalizace snadno odstraní.

I v případě této metody platí, že nalezení optimálních parametrů programu, tedy vlastně optimálního modelu telefonní linky, je možné pouze pomocí rozpoznávacích experimentů. Pouhým posuzováním upravených nahrávek je vhodné nastavení programu nalezitelné ještě obtížněji než v případě předchozí metody, na vhodný testovací signál pro identifikaci a na optimální velikost šumu z něj nelze usuzovat vůbec. Jediné, co lze zkoumáním výstupních nahrávek určit, je interval, v kterém by se zhruba měly pohybovat řády polynomů modelu. V případě polynomu A(z) jsem jako vhodné určil řády od 100. do 300., pro polynom B(z) se jevily jako vhodné řády od 10. do 30.

6.2.4 Popis konfigurace softwaru

Syntaxe pro spuštění této aplikace je ve tvaru:

```
MICtoTEL_im.exe <vstupní_soubor> <výstupní_soubor> <konfigurační_soubor>
```

Formáty vstupních souborů, které program podporuje, jsou naprosto shodné s formáty aplikace popsané v kapitole 6.1.4. Jako konfigurační soubor opět slouží běžný textový soubor, tentokrát je jeho obsah ve formátu:

```
normalizace 98
ekvalizace_zlomova_frekvence_filtru 3500
ekvalizace_delka_filtru 21
sum_velikost 25
soubor_koeficientu_polynomu_B c:\b.dat
soubor_koeficientu_polynomu_A c:\a.dat
```

I zde platí pravidlo, že chybějící řádek, neuvedená nebo nepřípustná hodnota parametru, vede k vyřazení dané funkce. Přehled jednotlivých parametrů a jejich přípustných hodnot obsahuje tabulka 6.6.

Tab. 6.6. Přehled parametrů a jejich přípustných hodnot

Označení parametru	Typ	Přípustné hodnoty
`normalizace`	reálné číslo	> 0
`ekvalizace_zlomova_frekvence_filtru`	celé číslo	> 0 ≤ 4000
`ekvalizace_delka_filtru`	celé liché číslo	> 0
`sum_velikost`	celé číslo	> 0
`soubor_koeficientu_polynomu_B`	řetězec	vše kromě mezery
`soubor_koeficientu_polynomu_A`	řetězec	vše kromě mezery

Struktura obou souborů s koeficienty polynomů je naprosto shodná, jedná se o běžné textové soubory, kde na každém řádku je jeden z koeficientů. První řádek odpovídá koeficientu nultému, s rostoucím pořadím řádků roste index koeficientů. Např. pro polynom 3.řádu $B(z) = b_0 + b_1 \cdot z^{-1} + b_2 \cdot z^{-2} + b_3 \cdot z^{-3}$ by soubor koeficientů v konkrétním případě $B(z) = 0{,}0164 - 0{,}0310 \cdot z^{-1} + 0{,}0180 \cdot z^{-2} - 0{,}0032 \cdot z^{-3}$ vypadal takto:

```
0.016366280826321
-0.031015061421715
0.018028967037737
-0.003173451696289
```

7 Rozpoznávací experimenty

Rozpoznávací experimenty jsou jediným objektivním nástrojem, pomocí kterého lze ověřit funkčnost a úspěšnost postupů a programů popsaných v předchozích kapitolách. Zároveň lze podle výsledků experimentů také určit konečné optimální nastavení simulačního softwaru. K samotnému provedení těchto testů je třeba nejprve zvolit konkrétní typ rozpoznávače, vybrat vhodný slovník slov pro rozpoznávání a pořídit nahrávky do trénovací a testovací množiny rozpoznávače.

Aby bylo možné z výsledků experimentů vyvodit nějaké závěry, je nejprve nutné provést dva základní testy. Při obou testech se použije rozpoznávač natrénovaný podle mikrofonních nahrávek. Jako testovací množina se v prvním případě použijí nahrávky pořízené rovněž mikrofonem, v druhém případě pomocí telefonu. U obou těchto testů se vyhodnotí úspěšnost rozpoznávání, tzv. rozpoznávací skóre, anglicky označované jako Correctness - správnost. Pro její výpočet platí vztah

$$\mathrm{Corr} = \frac{\mathrm{S}}{\mathrm{N}} \cdot 100 \quad [\%], \tag{7.1}$$

kde N je celkový počet rozpoznávaných slov a S je počet správně rozpoznaných slov. Lze očekávat, že úspěšnost druhého experimentu bude nižší vlivem různých vlastností trénovací a testovací množiny. Poté je třeba provést řadu rozpoznávacích experimentů, kdy jako testovací množina budou stále sloužit telefonní nahrávky a trénovací množina bude tvořena vždy různě upravenými mikrofonními nahrávkami. Pokud bude rozpoznávací skóre takových testů vyšší než v případě, kdy trénování probíhalo pomocí mikrofonních a testování pomocí telefonních nahrávek, lze konstatovat, že navržené metody simulace rušivých vlivů telefonní linky jsou funkční a splňují svůj účel. Čím optimálnější budou parametry simulačních programů, tím více se toto skóre bude blížit stavu, kdy byly pro trénování i testování použity nahrávky mikrofonní.

Rozpoznávací experimenty jsem prováděl na rozpoznávači izolovaných slov, jehož klasifikace je založena na použití skrytých markovských modelů (HMM) (viz. kapitola 5.4.2). Pro větší vypovídací hodnotu experimentů jsem zvolil slovník slov dříve používaný v reálné aplikaci, konkrétně v hlasovém dialogovém informačním systému InfoCity vyvinutém v Laboratoři počítačového zpracování řeči na Technické univerzitě v Liberci. Ten obsahuje celkem 218 slov, mezi kterými jsou např. názvy dnů v týdnu, některé číslovky, vybrané

autobusové zastávky a kulturní zařízení města Liberce, vybrané vlakové zastávky v České republice a klíčová slova pro ovládání aplikace.

7.1 Pořízení nahrávek trénovací a testovací množiny

Přestože trénovací množina, kterou jsem vytvořil, byla pořízena jen pro ověření funkčnosti metod simulace telefonní linky a od skutečné trénovací množiny použitelné v praxi se liší hlavně menším počtem realizací každého slova, snažil jsem se, aby splňovala všechny nároky na kvalitní testovací množinu kladené (viz. kapitola 3). Konkrétně jsem k pořizování nahrávek postupně přizval devět ženských a jedenáct mužských mluvčích ve věku od dvaceti do šedesáti let. Uvedená množina tedy obsahuje 20 různých realizací každého z 218 slov. Nahrávání jsem prováděl pomocí přenosného počítače, notebooku, se standardní zvukovou kartou a pomocí mikrofonu v podobě tzv. headsetu, tedy náhlavní soupravy, která v sobě kromě mikrofonu obsahuje také stereofonní sluchátka. S každým z mluvčích jsem postupně pořídil jednu několikaminutovou nahrávku obsahující všechna slova, v této nahrávce byly následně nalezeny hranice jednotlivých slov. Parametry každé z nahrávek byly: vzorkovací frekvence 8 kHz, 16 bitů na vzorek, kódování PCM.

Mnohem složitější situace nastala při pořizování nahrávek pro testovací množinu. Mikrofonní nahrávky jsem pořizoval shodným způsobem jako v případě trénovací množiny, k získání telefonních nahrávek jsem znovu použil telefonní kartu Dialogic D-21. Celkem jsem takto zaznamenal deset různých řečníků, přitom žádný z nich nebyl zároveň použit v trénovací množině. Aby měly rozpoznávací experimenty zmíněné na začátku kapitoly 7 určitou vypovídací hodnotu, bylo třeba mikrofonní i telefonní nahrávku každého mluvčího pořizovat současně. Jen tímto postupem bylo možné zajistit, aby realizace všech slov byly v případě mikrofonního i telefonního záznamu shodné a lišily se pouze svou technickou kvalitou. K tomu bylo třeba také nalézt shodný začátek obou nahrávek, aby hranice slov určené v jednom ze záznamů bylo možné bez úprav použít na záznam druhý. Toto hledání v různě kvalitních záznamech se ukázalo jako poměrně obtížné, ani metoda určení shodného začátku pomocí vzájemné korelace obou nahrávek nepřinesla příliš kvalitní výsledky. Nakonec se jako nejlepší a nejpřesnější řešení ukázalo hledat tyto začátky ručně pomocí běžného zvukového editoru. Nikdy ale samozřejmě nebylo možné dosáhnout přesnosti absolutní. Při znalosti funkce rozpoznávače, který signály zpracovává po framech délky 20 ms, což pro záznam se vzorkovací frekvencí 8 kHz představuje 160 vzorků, však lze tvrdit, že nepřesnost o velikosti několika málo vzorků je možné zanedbat.

Jako největší komplikace se při pořizování testovacích nahrávek ukázal problém nepřesnosti vzorkovacích frekvencí u zvukové a telefonní karty. Jeho vlivem docházelo k jevu, kdy se současně pořízené mikrofonní a telefonní nahrávky vzájemně rozcházely v čase. Odlišnost vzorkovacích frekvencí byla přitom tak velká, že mezi nahrávkami délky zhruba pět až osm minut (podle rychlosti řeči mluvčího) vznikal rozdíl až dvě sekundy. Ten bylo samozřejmě také nutné odstranit. Zvolil jsem k tomu postup, při kterém byl do kratšího záznamu vložen přesně takový počet vzorků, aby se délky obou záznamů sjednotily. Vzorky přitom nebyly do nahrávky vkládány s pravidelnou periodou, aby v prodlužované nahrávce nevzniklo vlivem těchto vzorků rušení na určité frekvenci. Odstup mezi dvěma vloženými vzorky byl proto vždy náhodně vybírán z určitého intervalu. Hodnota vkládaného vzorku byla jednoduše určena jako průměrná hodnota dvou vzorků, mezi které byl tento vzorek vložen, jednalo se tedy vlastně o lineární interpolaci. Takový způsob výpočtu je dostačující, stejným způsobem je také např. dopočítávána hodnota jednoho chybně načteného vzorku u přehrávačů kompaktních disků [3]. Druhým možným řešením, které se nabízelo, bylo nevkládat vzorky do kratšího záznamu, ale naopak vzorky vyjímat ze záznamu delšího. Tím by sice odpadla potřeba určovat hodnoty nových vzorků, ale upravovaný signál by bylo třeba ještě filtrovat vhodnou dolní propustí, aby nebyl porušen vzorkovací teorém (viz. kapitola 4.1), neboť vynecháním určitých vzorků vlastně dojde ke snížení vzorkovací frekvence.

7.2 Výsledky rozpoznávacích experimentů

Jak již bylo dříve zmíněno, veškeré rozpoznávací experimenty jsem prováděl na rozpoznávači izolovaných slov pracujícím se skrytými markovskými modely (HMM). Jeho konkrétní nastavení je zachyceno v tabulce 7.1.

Tab. 7.1. Nastavení rozpoznávače izolovaných slov

Počet stavů modelu	12
Počet mixtur stavu	1
Počet příznaků framu	18
Použité statické příznaky	8 LPCC kepstrálních koeficientů předzpracovaných metodou CMS
Použité dynamické příznaky	8 delta kepstrálních koeficientů, delta energie, delta delta energie

7.2.1 Experimenty s neupravenou trénovací množinou

Nejdříve jsem provedl dva testy s rozpoznávačem natrénovaným pomocí neupravené mikrofonní trénovací množiny. Z porovnání jejich výsledků s výsledky dalších experimentů bude možné následně usuzovat na úspěšnost navržených simulačních metod. V prvním experimentu jsem rozpoznávač testoval pomocí mikrofonní testovací množiny, druhý experiment jsem provedl s obsahově stejnou telefonní testovací množinou. Výsledné rozpoznávací skóre obou těchto testů určené podle vztahu (7.1) je uvedeno v tabulce 7.2.

Tab. 7.2. Výsledky experimentů prováděných s neupravenou trénovací množinou

Trénovací množina	**Testovací množina**	**Rozpoznávací skóre**
mikrofonní	mikrofonní	94,50 %
mikrofonní	telefonní	89,31 %

Výsledky obou testů potvrdily předpoklad, že úspěšnost druhého testu bude oproti testu prvnímu nižší. Je to dáno tím, že trénovací množina v tomto případě nesplňuje jeden ze základních požadavků kladených na takovou množinu, totiž požadavek na pořízení jejích nahrávek za stejných podmínek, tedy ve stejném prostředí a ve stejné technické kvalitě, v kterých bude rozpoznávač provozován.

Rozpoznávací skóre dosažené v prvním, čistě mikrofonním, testu svojí hodnotou 94,50 % není příliš dobré, v dnešních kvalitních rozpoznávačích izolovaných slov lze bez problémů dosáhnout úspěšnosti rozpoznávání kolem 97 až 98 %. Důvodem horšího výsledku je v tomto případě mnohem menší počet realizací každého slova v trénovací množině, která se od množin používaných v praxi značně liší právě tímto parametrem. Tato skutečnost však není pro určování funkčnosti a úspěšnosti navrženého simulačního softwaru důležitá, neboť veškeré závěry lze stanovit na základě rozdílů mezi výsledky jednotlivých experimentů a jejich absolutní hodnoty jsou z tohoto hlediska v podstatě nezajímavé.

Se znalostí výsledků uvedených v tabulce 7.2 jsem mohl uskutečnit další rozpoznávací experimenty s telefonní testovací množinou, při kterých mikrofonní trénovací množina byla upravena vždy jiným způsobem.

7.2.2 Experimenty s trénovací množinou upravenou bankou filtrů

Rozpoznávacích experimentů s trénovací množinou upravenou pomocí softwaru pracujícího s bankou filtrů jsem provedl celou řadu. Před každým testem jsem zmíněnou množinu upravoval softwarem s jinak nastavenými parametry. Tímto postupem jsem se snažil získat co možná nejlepší rozpoznávací skóre, tedy nalézt nejoptimálnější parametry tohoto simulačního softwaru. Měnitelnými parametry přitom byly hodnoty zlomových frekvencí a délek jednotlivých filtrů, velikosti šumu a brumu. Právě na hledání vhodných velikostí šumu a brumu jsem se zaměřil nejvíce. Pro počáteční experimenty jsem jako výchozí použil hodnoty parametrů uvedených v tabulkách 6.3 a 6.4, z jejich výsledků jsem potom volil nové hodnoty parametrů pro další testy. Po celou dobu experimentů jsem pracoval odděleně s parametry programu pro pevnou a mobilní telefonní síť. Až na několik výjimek jsem většinu experimentů provedl se smíšenou trénovací množinou, ve které byla jedna polovina záznamů upravena na pevnou telefonní síť a druhá na síť mobilní. V některých testech jsem také pracoval s nastavením simulačního softwaru na průměrnou síť, pro kterou jsem veškeré parametry určil jako aritmetické průměry z hodnot parametrů pevné a mobilní telefonní sítě. Nejzajímavější výsledky experimentů jsou shrnuty v následujících tabulkách a grafech.

Tabulka 7.3 zobrazuje výsledky některých experimentů, které jsem provedl s úmyslem zjistit, zda je vhodnější trénovací množinu upravovat pouze na vlastnosti jedné telefonní sítě, pevné či mobilní, nebo zda je lepší při úpravě uvažovat obě uvedené sítě. Úprava trénovací množiny označená „pevná + mobilní síť“ vznikla již zmíněným způsobem, kdy polovina záznamů byla upravena na pevnou síť a polovina na síť mobilní. Veškeré experimenty z tabulky 7.3 jsem provedl s parametry dolní a horní propusti podle tabulky 6.3. Z výsledků těchto experimentů je zřejmé, že v naprosté většině případů je vhodné do trénovací množiny zahrnout vlastnosti obou dvou typů telefonních sítí. Jako vhodnější prostředek reprezentace vlastností obou sítí se přitom jeví smíšená trénovací množina, která oproti množině průměrné přináší lepší výsledky.

Vliv ekvalizace a kompenzace účinku dolní propusti na rozpoznávací skóre demonstruje tabulka 7.4. Veškeré testy v ní uvedené jsem provedl se smíšenou trénovací množinou s již zmíněným nastavením (viz. tabulka 6.3). Podle dosažených výsledků lze konstatovat, že zařazení již zmíněných úprav signálu do simulačního softwaru je opodstatněné, neboť téměř vždy vede ke zlepšení úspěšnosti rozpoznávání.

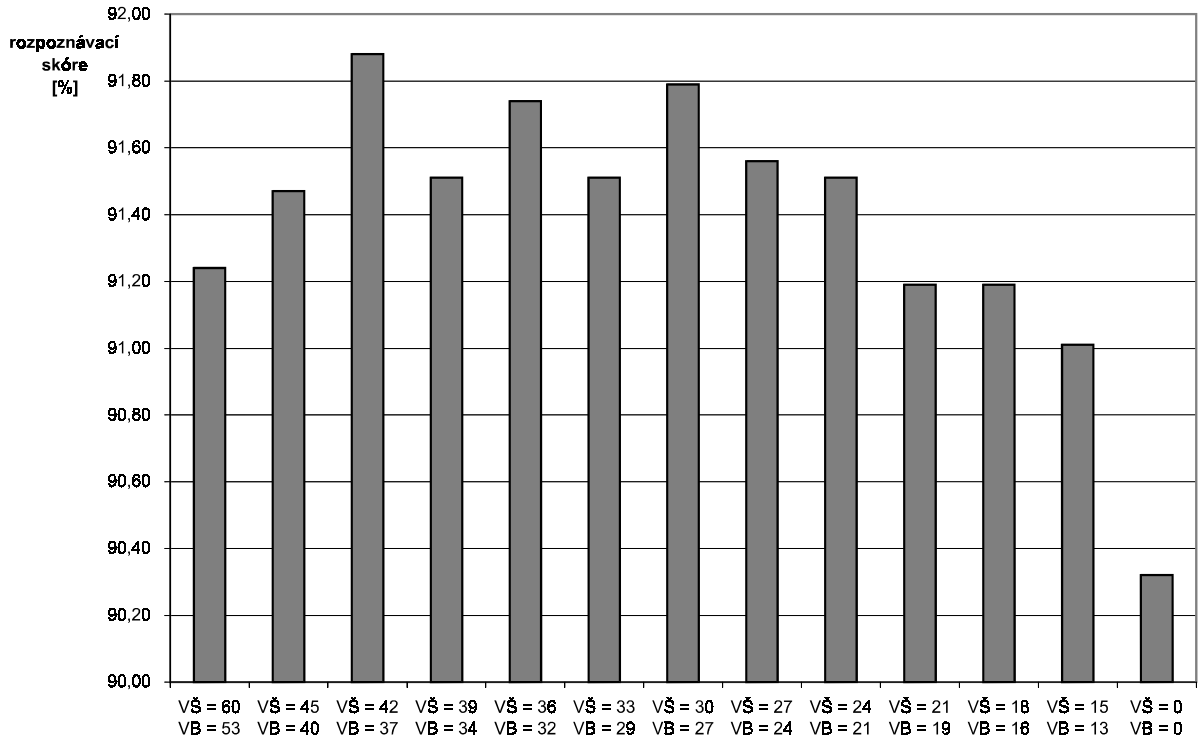
Tab. 7.3. Porovnání výsledků experimentů při různých úpravách trénovací množiny

	Úprava trénovací množiny	Rozpoznávací skóre
Velikost šumu = 60 **Velikost brumu = 53**	**pevná + mobilní síť**	91,24 %
	průměrná síť	90,92 %
	pevná síť	91,28 %
	mobilní síť	90,64 %
Velikost šumu = 42 **Velikost brumu = 37**	**pevná + mobilní síť**	91,88 %
	průměrná síť	91,65 %
	pevná síť	91,47 %
	mobilní síť	91,65 %
Velikost šumu = 30 **Velikost brumu = 0**	**pevná + mobilní síť**	92,02 %
	průměrná síť	91,74 %
	pevná síť	91,47 %
	mobilní síť	91,56 %
Velikost šumu = 0 **Velikost brumu = 0**	**pevná + mobilní síť**	90,32 %
	průměrná síť	90,37 %
	pevná síť	90,41 %
	mobilní síť	90,05 %

Tab. 7.4. Výsledky experimentů demonstrujících vliv ekvalizace a kompenzace účinku DP

		Rozpoznávací skóre
Velikost šumu = 60 **Velikost brumu = 53**	**ekvalizace + kompenzace DP**	91,24 %
	ekvalizace vyřazena	90,32 %
	kompenzace DP vyřazena	90,92 %
Velikost šumu = 30 **Velikost brumu = 27**	**ekvalizace + kompenzace DP**	91,79 %
	ekvalizace vyřazena	91,24 %
	kompenzace DP vyřazena	91,79 %
Velikost šumu = 0 **Velikost brumu = 0**	**ekvalizace + kompenzace DP**	90,32 %
	ekvalizace vyřazena	90,46 %
	kompenzace DP vyřazena	88,85 %

Úspěšnost rozpoznávání se smíšenou trénovací množinou (dle tabulky 6.3) při různých velikostech šumu a brumu zobrazuje obrázek 7.1. Nejlepších výsledků přitom bylo dosaženo při hodnotách šumu 42, 36 a 30, tedy hodnotách nižších, než je výchozí hodnota zjištěná analýzou nahrávek (viz. tabulka 6.4). Tento rozpor je zřejmě dán faktem, že velikost šumu a brumu jsem pomocí analýzy záznamů určoval pouze z úseků mezi jednotlivými slovy. V těchto úsecích přitom mohl být vlivem automatického řízení úrovně zesílení na telefonní lince šum větší oproti šumu obsaženém ve slovech. Pro zmíněné tři hodnoty šumu jsem následně provedl další experimenty s různými velikostmi brumu.



Obr. 7.1. Vliv velikosti šumu (VŠ) a velikosti brumu (VB) na úspěšnost rozpoznávání

Pokusů se změnou vlastností dolní a horní propusti jsem provedl pouze několik, a to na parametrech průměrné telefonní sítě (dle tabulky 6.3) při velikosti šumu 30 a velikosti brumu 0. Konkrétně jsem postupně zužoval a rozšiřoval propustná pásma a měnil strmosti filtrů. Jak však zobrazuje tabulka 7.5, žádná z těchto změn nevedla ke zlepšení rozpoznávacího skóre.

Tab. 7.5. Vliv změny vlastností DP a HP na výsledky experimentů

Úprava parametrů dolní a horní propusti	**Rozpoznávací skóre**
bez úpravy parametrů	91,74 %
zúžení propustných pásem	90,55 %
rozšíření propustných pásem	91,24 %
snížení strmosti filtrů	91,74 %
zvýšení strmosti filtrů	91,70 %

Mimo již uvedených jsem provedl ještě celou řadu dalších experimentů, celkem jich bylo zhruba sedmdesát. Nejlepšího rozpoznávacího skóre 92,02 % jsem dosáhl při použití smíšené trénovací množiny s nastavením simulačního programu uvedeným v tabulkách 7.6 a 7.7.

Tab. 7.6. Optimální parametry různé pro pevnou a mobilní telefonní síť

		Pevná telefonní síť	**Mobilní telefonní síť**
Dolní propust	**zlomová frekvence**	3550 Hz	3480 Hz
	délka filtru	41 vzorků	45 vzorků
Horní propust	**zlomová frekvence**	200 Hz	225 Hz
	délka filtru	181 vzorků	151 vzorků

Tab. 7.7. Optimální parametry shodné pro pevnou a mobilní telefonní síť

Normalizace	**na hodnotu**	98 %
Ekvalizace	**zlomová frekvence filtru**	3500 Hz
	délka filtru	21 vzorků
Dolní propust	**kompenzace účinku**	2,5 %
Šum	**velikost**	30
Brum	**velikost**	0
	zlomová frekvence filtru	50 Hz
	délka filtru	65 vzorků

7.2.3 Experimenty s trénovací množinou upravenou identifikovaným modelem

Také pro nalezení optimálních parametrů simulačního softwaru pracujícího s identifikovaným modelem jsem musel provést značný počet rozpoznávacích experimentů s různě upravenou trénovací množinou. Testy jsem postupně prováděl se všemi testovacími signály, u každého z nich jsem hledal optimální řády polynomů A(z) a B(z) a optimální velikost šumu. Při volbě řádů polynomů modelu jsem se pohyboval v intervalech zmíněných v kapitole 6.2.3.

Tabulka 7.8 zachycuje výsledky počátečních experimentů, tedy parametry nejlepšího ARX modelu pro každý použitelný testovací signál. Velikost šumu jsem při těchto experimentech zvolil jako nulovou, polynom A(z) jsem použil 200. řádu, hodnoty řádů polynomu B(z) jsem volil z intervalu 10 až 20. Pro úspěšnější identifikační signály jsem následně provedl mnohem podrobnější testy.

Tab. 7.8. Výsledky počátečních experimentů pro různé modely ARX

Testovací signál identifikace	Řád A(z)	Řád B(z)	Rozpoznávací skóre
rozmítaný signál – 1. verze	200.	15.	82,98 %
rozmítaný signál – 2. verze	200.	15.	84,13 %
mluvená řeč – 1. verze	200.	15.	87,34 %
mluvená řeč – 2. verze	200.	20.	88,49 %
mluvená řeč – 3. verze	200.	15.	88,67 %
jednotkové skoky – 1. verze	200.	20.	90,96 %
jednotkové skoky – 2. verze	200.	10.	89,86 %
Bílý šum – 1. verze	200.	20.	87,94 %

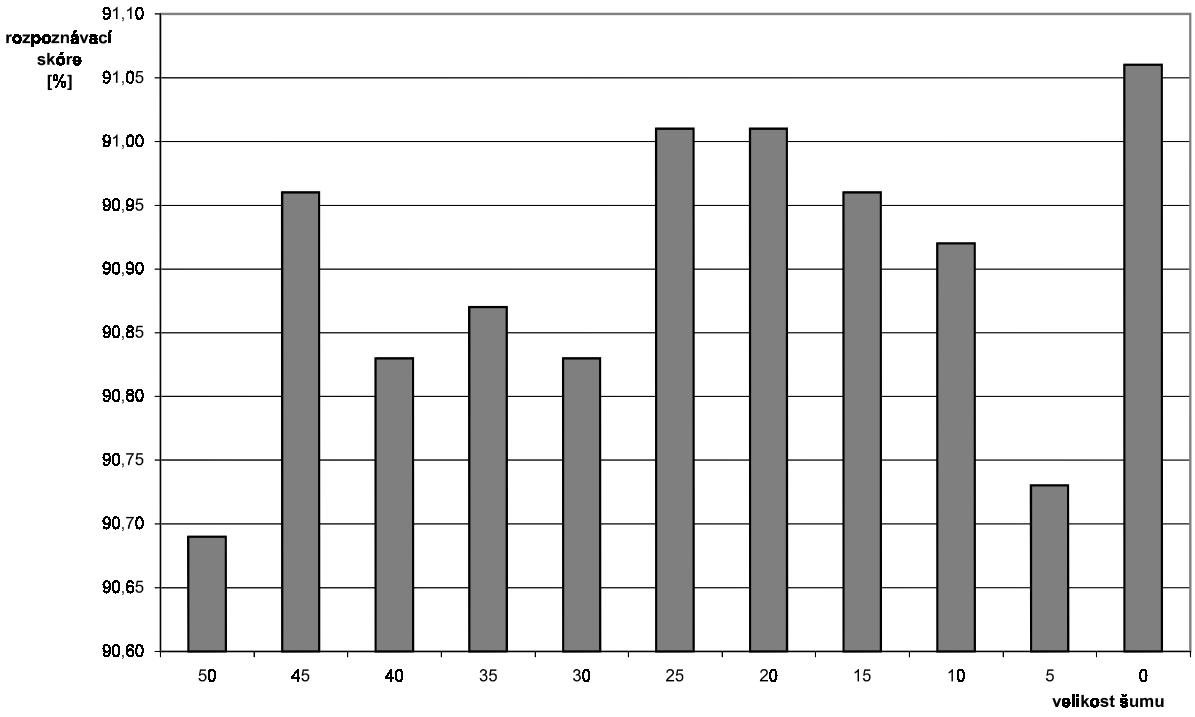
Cílem podrobnějších testů bylo nalezení optimálních řádů polynomů A(z) a B(z) pro dané testovací signály. Ukázku výsledků takových testů zobrazuje tabulka 7.9, ve které jsou uvedeny konkrétně výsledky některých experimentů při nulovém šumu s modely odhadnutými z jednotkových skoků (1. verze). Ze zmíněné tabulky je také dobře patrné, jak volba řádů polynomů modelu ovlivňuje výslednou úspěšnost rozpoznávání.

Závislost úspěšnosti rozpoznávání na zvolené velikosti šumu ukazuje obrázek 7.2. Všechny v něm zobrazené experimenty byly prováděny s mikrofonní trénovací množinou upravenou pomocí ARX modelu, jehož parametry byly odhadnuty pomocí testovacího signálu s jednotkovými skoky (1. verze), s polynomem A(z) 150. řádu a polynomem B(z) 20. řádu.

Z výsledků vyplývá, že přijatelná velikost šumu se pohybuje kolem hodnoty 20, nejlepších výsledků však bylo (i u jiných modelů) dosaženo se šumem nulovým.

Tab. 7.9. Demonstrace závislosti rozpoznávacího skóre na volbě řádů polynomů

Řád polynomu A(z)	Řád polynomu B(z)	Rozpoznávací skóre
125.	20.	90,87 %
145.	20.	90,78 %
150.	16.	90,32 %
150.	20.	91,06 %
150.	24.	88,81 %
155.	20.	90,78 %
200.	15.	89,63 %
200.	20.	90,96 %
200.	25.	88,26 %
300.	20.	90,64 %



Obr. 7.2. Vliv velikosti šumu na úspěšnost rozpoznávání

Stejně jako v případě první metody i s identifikovanými modely jsem celkem provedl něco přes sedmdesát experimentů. Nejvyšší úspěšnosti rozpoznávání 91,06 % jsem přitom dosáhl s modelem, jehož parametry byly odhadnuty z první verze testovacího signálu s jednotkovými skoky. Přesné optimální nastavení simulačního softwaru je uvedeno v tabulce 7.10.

Tab. 7.10. Optimální parametry simulačního softwaru pracujícího s identifikovaným modelem

Normalizace	**na hodnotu**	98 %
Ekvalizace	**zlomová frekvence filtru**	3500 Hz
	délka filtru	21 vzorků
Řád polynomu	**čitatele B(z)**	20.
	jmenovatele A(z)	150.
Šum	**velikost**	0

7.2.4 Experimenty s trénovací množinou upravenou kombinací obou metod

Po nalezení nejoptimálnějších parametrů obou metod jsem provedl několik experimentů, při kterých byly tyto metody různým způsobem kombinovány. Jako jedna z kombinací se nabízela možnost upravit jednu polovinu mikrofonní trénovací množiny pomocí první metody, druhou polovinu pomocí metody druhé (1. způsob). Také jsem vyzkoušel celou trénovací množinu přizpůsobit telefonní lince pomocí obou metod současně, tedy nejprve úpravu nahrávek provést jedním programem a hned na to na již upravené nahrávky aplikovat také druhý program. Použití softwaru pracujícího nejprve s bankou filtrů a následně s identifikovaným modelem je označeno jako 2. způsob, opačný postup jako 3. způsob. Výsledky všech těchto experimentů, při kterých byly použity optimální parametry obou simulačních programů z tabulek 7.6, 7.7 a 7.10, zachycuje tabulka 7.11.

Tab. 7.11. Porovnání úspěšnosti rozpoznávání při různých kombinacích obou metod

Úprava mikrofonní trénovací množiny	**Rozpoznávací skóre**
1. způsob – banka filtrů + identifikovaný model	91,47 %
2. způsob – banka filtrů, identifikovaný model	90,73 %
3. způsob – identifikovaný model, banka filtrů	91,61 %

7.3 Zhodnocení výsledků rozpoznávacích experimentů

Přehled výsledků rozpoznávacích experimentů, z kterých lze usuzovat na úspěšnost navržených simulačních metod telefonní linky, je uveden v tabulce 7.12. Ta konkrétně obsahuje rozpoznávací skóre celkem pěti experimentů. První dva byly provedeny s neupravenou mikrofonní trénovací množinou, v prvním experimentu se k testování použily mikrofonní nahrávky, v druhém nahrávky telefonní. U dalších tří testů, před kterými byla mikrofonní trénovací množina upravena vždy jinou metodou pomocí softwaru nakonfigurovaného na nejlepší známé parametry, je v tabulce navíc uvedeno tzv. poměrné zlepšení rozpoznávacího skóre. To vlastně vyjadřuje, o kolik procent se zlepšila úspěšnost rozpoznávání telefonních záznamů po úpravě trénovací množiny v poměru ke zhoršení úspěšnosti při přechodu od rozpoznávání mikrofonních nahrávek k nahrávkám telefonním s neupravenou trénovací množinou. Vztah pro výpočet poměrného zlepšení rozpoznávacího skóre po úpravě trénovací množiny určitou metodou lze tedy vyjádřit jako

$$\mathrm{PzCorr} = \frac{\mathrm{Corr_{UMT}} - \mathrm{Corr_{MT}}}{\mathrm{Corr_{MM}} - \mathrm{Corr_{MT}}} \cdot 100 \quad [\%], \tag{7.2}$$

kde $\mathrm{Corr_{UMT}}$ je rozpoznávací skóre telefonních nahrávek při upravené mikrofonní trénovací množině, $\mathrm{Corr_{MT}}$ je rozpoznávací skóre telefonních nahrávek při neupravené mikrofonní trénovací množině a $\mathrm{Corr_{MM}}$ je rozpoznávací skóre mikrofonních nahrávek při neupravené mikrofonní trénovací množině.

Tab. 7.12. Porovnání úspěšnosti jednotlivých metod simulace telefonní linky

Trénovací množina	**Testovací množina**	**Rozpoznávací skóre**	**Poměrné zlepšení rozpoznávacího skóre**
mikrofonní neupravená	mikrofonní	94,50 %	-
mikrofonní neupravená	telefonní	89,31 %	-
mikrofonní upravená bankou filtrů	telefonní	92,02 %	52,22 %
mikrofonní upravená identifikovaným modelem	telefonní	91,06 %	33,72 %
mikrofonní upravená kombinací obou metod	telefonní	91,61 %	44,32 %

Je celkem logické, že vhodnější by bylo úspěšnost simulačních metod nevztahovat k výsledkům rozpoznávání mikrofonních nahrávek rozpoznávačem s mikrofonní trénovací množinou, ale k výsledkům rozpoznávání telefonních nahrávek rozpoznávačem se skutečnou telefonní trénovací množinou. V tom případě by vztah (7.2) měl podobu

$$\mathrm{PzCorr} = \frac{\mathrm{Corr_{UMT}} - \mathrm{Corr_{MT}}}{\mathrm{Corr_{TT}} - \mathrm{Corr_{MT}}} \cdot 100 \quad [\%], \tag{7.3}$$

kde $\mathrm{Corr_{TT}}$ je již zmíněné rozpoznávací skóre telefonních nahrávek při telefonní trénovací množině. Pokud by tedy úspěšnost rozpoznávání telefonních nahrávek byla shodná pro rozpoznávač natrénovaný podle telefonní i podle upravené mikrofonní množiny, hodnota vztahu (7.3) by byla 100 %. Naproti tomu vztah (7.2) by mohl nabýt hodnoty 100 % pouze v případě, že rozpoznávač s upravenou trénovací množinou by při rozpoznávání telefonních nahrávek dosáhl stejných výsledků jako rozpoznávač natrénovaný na mikrofonních nahrávkách při rozpoznávání záznamů pořízených mikrofonem. Toho lze ovšem jen velmi těžko dosáhnout, neboť telefonní linka do značné míry akustický signál ovlivňuje, a tím může dojít ke zkreslení nebo dokonce ztrátě některých charakteristik lidské řeči používaných pro její rozpoznávání. I přes to, že poměrné zlepšení rozpoznávacího skóre určené podle vztahu (7.3) má mnohem větší vypovídací hodnotu o kvalitách simulačních metod, pro výpočet hodnot v tabulce 7.12 jsem použil vzorce (7.2), neboť kvůli technickým omezením na počítačové telefonní kartě nebylo možné pořídit potřebnou telefonní trénovací množinu.

Z výsledků uvedených v tabulce 7.12 vyplývá, že navržené metody a programy pro úpravu mikrofonních nahrávek na telefonní plní svůj účel, tedy zlepšují rozpoznávání řeči na telefonní lince bez nutnosti pořizovat pro rozpoznávač telefonní trénovací množinu. Jako úspěšnější se nakonec ukázala metoda využívající banku filtrů. Důvodem toho je zřejmě jednodušší a intuitivnější postup pro hledání optimálních parametrů této metody. Metoda pracující s identifikovaným modelem byla navíc znevýhodněna již jednou zmíněným technickým omezením, při kterém bylo možné pořizovat nahrávky s testovacími identifikačními signály pouze v rámci jedné pobočkové ústředny, zatímco nahrávky analyzované pro potřeby metody využívající banku filtrů byly mnohem rozmanitější. Při rozpoznávacích experimentech bylo také zjištěno, že kombinace obou simulačních metod již nepřináší žádné další zlepšení. Naopak jako velmi výhodné se ukázalo uvažovat rozdíly mezi pevnou a mobilní telefonní sítí, tedy určovat, pokud je to možné, parametry pro každou síť odděleně.

8 Závěr

Hlavním cílem diplomové práce bylo navrhnout metody simulace rušivých vlivů přenosové cesty telefonní linky, které by po aplikaci na mikrofonní trénovací množinu rozpoznávače vedly ke zvýšení úspěšnosti rozpoznávání řeči po telefonu bez nutnosti pořizovat skutečnou telefonní trénovací množinu. Další úkol spočíval v realizaci těchto metod v kompaktním programu s možností vnější konfigurace, v ověření jejich funkčnosti a v nalezení optimálních parametrů zmíněného programu.

Po seznámení s problematikou rozpoznávání řeči a po prozkoumání vlastností telefonních sítí jsem vytvořil dvě principiálně odlišné metody pro simulaci telefonní linky. Zatímco jedna z metod se snaží popsat jednotlivé rušivé vlivy této linky odděleně, druhá nahlíží na celou přenosovou cestu telefonu jako na jeden celek. Vzhledem k tomu, že obě metody jsou poměrně rozdílné, vytvořil jsem pro každou z nich samostatný simulační program. Oba tyto programy typu konzolová aplikace jsou pro uživatele jednoduše konfigurovatelné pomocí textového souboru. Díky tomu, že záznamy trénovací množiny jsou upravovány ještě před samotným vstupem do rozpoznávače, není použití programů vázáno jen na jeden konkrétní rozpoznávač, ale jsou použitelné s libovolným systémem pro rozpoznávání, který používá záznam zvuku ve formátu 8 kHz, 16 bitů na vzorek, kódování PCM. Tento formát nebyl zvolen náhodně, jedná se o jakýsi standard používaný pro rozpoznávání mluvené řeči, navíc je schopen pojmout celé telefonní pásmo, tedy také frekvenční rozsah výstupních nahrávek simulace, bez ztráty informace. Zvukové záznamy jiných formátů by musely být pro potřeby simulace telefonní linky do zmíněného formátu převedeny.

Pomocí simulací a rozpoznávacích experimentů jsem nalezl optimální parametry každého z programů a následně ověřil dalšími rozpoznávacími experimenty jejich funkčnost. Lze konstatovat, že oba simulační přístupy plní svůj účel, neboť zlepšují výsledky rozpoznávání řeči po telefonu, pokud jsou aplikovány na trénovací množinu řečového rozpoznávače. Konkrétní výsledky těchto experimentů jsou uvedeny v tabulce 7.12.

Do budoucna by bylo vhodné provádět případné další srovnávací experimenty také se skutečnou telefonní trénovací množinou, aby z jejich výsledků bylo možné objektivněji určit úspěšnost jednotlivých simulačních metod. Tyto experimenty s telefonní trénovací množinou se mi bohužel nepodařilo kvůli již zmíněným technickým omezením realizovat.

Výsledky metody pracující s identifikovaným modelem by bylo možné zřejmě ještě zvýšit, pokud by se pořízení testovacích signálů pro identifikaci provedlo na lince vhodnější, než je linka pobočkové ústředny, tedy ve skutečné pevné či mobilní telefonní síti. Technické podmínky mi bohužel neumožňovaly takové nahrávky pořídit.

Jistě by bylo přínosem také ověřit, jak dalece jsou nalezené optimální parametry svázány s konkrétním typem použité telefonní karty, tedy určit, zda je potřebné pro každou další telefonní kartu hledat parametry nové.

9 Literatura

[1] Psutka, J.: Komunikace s počítačem mluvenou řečí. Academia, Praha 1995

[2] Nouza, J. a kol.: Počítačové zpracování řeči - cíle, problémy, metody a aplikace. Sborník článků. Technická univerzita v Liberci, Liberec 2001

[3] Salava, T. a kol.: Přehrávače číslicových zvukových desek systému CD. SNTL, Praha 1991

[4] Nouza, J.: Počítačové zpracování signálů. Studijní materiály. Technická univerzita v Liberci, Liberec 2003
http://www.fm.vslib.cz/~kes/pages/nove_predmety/pzs/ramce_main.html

[5] Skalický, P.: Digitální filtrace a signálové procesory. Skripta ČVUT, Praha 1997

[6] Bellman, R. E.: Dynamic Programming. Princeton University Press, Princeton NJ 1957

[7] Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of The IEEE, vol. 77, no. 2, 1989

[8] Modrlák, O.: Teorie automatického řízení II - Úvod do diskrétní parametrické identifikace. Studijní materiály. Technická univerzita v Liberci, Liberec 2002
http://www.fm.vslib.cz/~krtsub/fm/tr2/tar2_did.pdf

[9] Ljung, L.: System Identification Toolbox For Use with MATLAB. The MathWorks, Inc., Natick 2002

[10] Holada, M., Nouza, J.: Searching for Methods and Parameters for More Reliable Recognition of Telephone Speech. Proc. of Radioelektronika '98, Brno 1998

[11] Weintraub, M., Neumeyer, L.: Constructing Telephone Acoustic Models from a High-Quality Speech Corpus. Proc. of The 1994 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '94), Adelaide 1994

[12] Neumeyer, L., Digalakis, V., Weintraub, M.: Training Issues and Chanel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus. IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, 1994

[13] Semenec, P., Holada, M.: The Acoustic Model of Phone Line for ASR Databases Recorded by Microphone. Proc. of Radioelektronika 2003, Brno 2003

10 Přílohy

10.1 Slovník slov použitých v rozpoznávacích experimentech

ano	Malé	Ostašov	Broumov
ne	Lípa	Kryštofovo	Brno
pondělí	Varšava	Údolí	Cvikov
úterý	Dům kultury	Kunratická	Česká Kamenice
středa	P.K.O.	Fignerka	Česká Lípa
čtvrtek	Golet	Bedřichov	Děčín
pátek	Hanychov	Bílý Kostel	Dolní
sobota	Lidové Sady	Český Dub	Poustevna
neděle	Fígnerova	Fojtka	Dvůr Králové
dnes	Vratislavice	Frýdlant	Harrachov
zítra	Pekárny	v Čechách	Hlinsko
doprava	Pavlovice	Hodkovice	Hořice
městská	Letná	nad Mohelkou	Hradec Králové
autobusová	Kateřinky	Horní Vítkov	Chomutov
železniční	Doubí	Hrádek	Chrudim
místní	Ruprechtice	nad Nisou	Jablonné
dálkové	Harcov	Chrastava	v Podještědí
mezinárodní	Rudolfov	Jablonec	Jaroměř
banky	Rochlice	nad Nisou	Jičín
spořitelna	Pilinkov	Libverda	Jilemnice
komerční	Šimonovice	Mníšek	Kadaň
obchodní	Králův Háj	Nové Město	Karlovy Vary
investiční	České Mládeže	pod Smrkem	Klášterec
í pé bé	Novoplastik	Oldřichov	nad Ohří
obchody	Růžodol	v Hájích	Kolín
Tesco	Broumovská	Osečná	Litoměřice
knihovna	Krásná	Raspenava	Litomyšl
bazén	Studánka	Václavice	Lomnice
kultura	Lites	Frýdlant	nad Popelkou
divadla	Radčice	Hodkovice	Louny
kina	Elitex	Hrádek	Lovosice
muzea	Nádraží	Jablonec	Mladá Boleslav
galerie	Zelené Údolí	Nové Město	Mnichovo
Šaldovo	Vesec	Oldřichov	Hradiště
Naivní	Lukášov	Bílina	Most

Náchod	Žatec	nápověda	jedenáct
Nová Paka	Železný Brod	návrat	dvanáct
Nový Bor	Berlín	nevím	třináct
Nymburk	Drážďany	opakuj	čtrnáct
Pardubice	Jelenia Gora	ostatní	patnáct
Pec	Zgorzelec	podrobnosti	šestnáct
pod Sněžkou	Žitava	předchozí	sedmnáct
Podbořany	sport	vyjmenuj	osmnáct
Poděbrady	florbal	zpátky	devatenáct
Praha	fotbal	všechny	dvacet
Rumburk	hokej	dopoledne	dvacet jedna
Semily	atletika	odpoledne	dvacet dva
Svitavy	tenis	nula	dvacet tři
Šluknov	volejbal	jedna	dvacet čtyři
Špindlerův	košíková	dva	dvacet pět
Mlýn	turistika	dvě	dvacet šest
Tanvald	lyžování	tři	dvacet sedm
Teplice	kuželky	čtyři	dvacet osm
Trutnov	další	pět	dvacet devět
Turnov	děkuji	šest	třicet
Ústí nad Labem	dozadu	sedm	třicet jedna
Varnsdorf	chyba	osm	
Vrchlabí	konec	devět	
Vysoké Mýto	končit	deset	

10.2 Datová příloha na disku CD-R

Součástí diplomové práce je disk CD-R, který obsahuje elektronickou verzi tohoto dokumentu ve formátu PDF, oba simulační programy pracující jak s bankou filtrů, tak s identifikovaným modelem (spustitelné soubory i projekty pro Microsoft Visual C++ 6.0 se zdrojovým kódem), optimální konfiguraci každého z programů a trénovací a testovací množiny použité při rozpoznávacích experimentech. Podrobnější popis obsahu CD je uveden v HTML souboru `Index.htm`, který se díky funkci Autorun otevře automaticky po vložení disku do mechaniky. Pokud by k tomu z nějakého důvodu nedošlo, lze jej na CD nalézt v jeho kořenovém adresáři.